



**DELHI UNIVERSITY
LIBRARY**

DELHI UNIVERSITY LIBRARY

CI No. T15:87

H6: 1

Ac. No. 173255

Date of release for loan

This book should be returned on or before the date last stamped below. An overdue charge of 06 nP will be charged for each day the book is kept overtime.

[illegible]

**MEASUREMENT AND EVALUATION
IN THE
ELEMENTARY SCHOOL**

MEASUREMENT AND EVALUATION IN THE ELEMENTARY SCHOOL

BY

HARRY A. GREENE, PH.D.

PROFESSOR OF EDUCATION AND
DIRECTOR OF BUREAU OF EDUCATIONAL RESEARCH AND SERVICE
UNIVERSITY OF IOWA

ALBERT N. JORGENSEN, PH.D.

PRESIDENT
UNIVERSITY OF CONNECTICUT

and

J. RAYMOND GERBERICH, PH.D.

ASSOCIATE PROFESSOR OF EDUCATION AND
DIRECTOR OF BUREAU OF EDUCATIONAL RESEARCH AND SERVICE
UNIVERSITY OF CONNECTICUT

LONGMANS, GREEN AND CO.

NEW YORK • LONDON • TORONTO

LONGMANS, GREEN AND CO, INC.
55 FIFTH AVENUE, NEW YORK 3

LONGMANS, GREEN AND CO, LTD.
OF PATERNOSTER ROW
43 ALBERT DRIVE, LONDON, S W 19
17 CHITTARANJAN AVENUE, CALCUTTA 13
NICOL ROAD, BOMBAY 1
36A MOUNT ROAD, MADRAS 2

LONGMANS, GREEN AND CO.
215 VICTORIA STREET, TORONTO 1

GREENE, JORGENSEN & GERBERICH 1 1
MEASUREMENT AND EVALUATION IN THE
ELEMENTARY SCHOOL

COPYRIGHT • 1942
BY LONGMANS, GREEN AND CO, INC

ALL RIGHTS RESERVED, INCLUDING THE RIGHT TO REPRODUCE
THIS BOOK, OR ANY PORTION THEREOF, IN ANY FORM

PUBLISHED SIMULTANEOUSLY IN
THE DOMINION OF CANADA BY
LONGMANS, GREEN AND CO, TORONTO

First edition January 1942
Reprinted August 1943
Reprinted October 1945
July 1946, ~~July~~ 1947

Printed in the United States of America
VAN REES PRESS • NEW YORK

PREFACE

This book is designed especially for the use of elementary school teachers and students of elementary education. It presents a practical introductory discussion of the essential principles of measurement and evaluation in the elementary school. This volume is essentially a completely revised and expanded treatment of an earlier volume which appeared under the title of *The Use and Interpretation of Elementary School Tests* in 1935. The 1935 edition of the book was itself a revision of a general text entitled *The Use and Interpretation of Educational Tests* which first appeared in 1929. Just as this first general volume later seemed to be somewhat unsuited for the use of both elementary and secondary school teachers, certain inadequacies in the illustrative materials, recent significant changes in point of view and in methods and techniques of measurement since the appearance of the 1935 edition have served to make prudent this further revision.

Students and teachers whose major interests are in the elementary school represent the group specifically addressed in this volume. A second volume, parallel in general organization and treatment, is just as specifically addressed to those teachers and students who primarily face the problems of instruction, measurement, and evaluation at the secondary school level. In each volume, the illustrations, examples, and problems are chosen from material designed for use in the grades under consideration. Many of the problems of measurement are common to both the secondary and elementary levels. However, illustrations are more meaningful if taken from fields close to the interest of the students and teachers. The revision of these two volumes succeeds in bringing this treatment of measurements and evaluation in elementary education quite up to the best thought and practices of 1942.

The decade just past has been marked by many important developments in curricular points of view, in instructional methods, in measurement and evaluation techniques. In

this revision of the elementary volume a special effort has been made to broaden the point of view reflected by the authors in their earlier treatments and to introduce the student to an easily comprehended discussion of the newest and best evaluation techniques which have thus far appeared. In this treatment an increased emphasis is given to the methods and materials designed for the measurement of intelligence and the evaluation of certain of the more intangible aspects of the child's personality. Many of the instruments and procedures presented here are new-comers to the field of measurement. Some have only recently proved their dependable worth for the educational practitioner.

In this volume the authors have placed a heavy stress on the crucial problems of improving the teacher-made examination and test. By principle and by example the construction, use, improvement, and interpretation of all types of evaluative and measuring instruments are treated in detail. Completely new material is presented on measurement of personality, physical and health education. The simplified treatment of the statistical problems related to measurement and evaluation presented in the earlier editions of these books is continued in this volume. Completely new problems dealing with the interpretation of test results closely related to the elementary school teacher's actual needs have been prepared. A completely new revision of the work-book to accompany this text is also available.

This revised volume is planned to provide a complete and systematic handbook for any student or teacher requiring a straightforward discussion of about all of the fundamental ideas and techniques of evaluation in elementary education. It is written from the point of view of the classroom teacher. At all possible points over-technical language is avoided. In instances in which technical language cannot be avoided, such terms are introduced in context, defined, and illustrated. Many words which may lie outside of the experience of the reader are included in the *Glossary*.

To the classroom teacher and the supervisor of the elementary grades, as well as to the normal school and teachers college student, this volume offers carefully selected suggestions of ways in which measurement and evaluation instru-

ments may be effectively used in the teaching of children. In addition, many general hints are given for the guidance of the student and teacher in constructing, selecting, using, and interpreting educational tests as valuable aids in accomplishing this task.

Grateful acknowledgment is here expressed to the many experienced teachers and supervisors who have contributed directly and indirectly to the formulation of much of the material incorporated into this volume. The authors are especially indebted to the many users of the earlier editions of this text who by their friendly and critical comments have stimulated the development of this volume in its present form.

EDITOR'S INTRODUCTION

Previous editions of this book have proven dependable and welcome instruments of instruction and guidance in the field of educational tests and measurements. For years young teachers have obtained their first grasp of the problems and possibilities of measurement through a study of the carefully and intelligently written pages of this book. Principals and classroom teachers have used it in planning and prosecuting the actual measurement of actual children in a great variety of ways.

Even the most successful book requires occasional revision. In education, as well as in other fields, new information becomes available, new points of view appear, a new assortment of emphases here and there becomes wise, and old editions of books become less and less useful. The authors of this text are wise in doing two things in the present edition: (1) maintaining the effective presentation of the whole problem of educational measurement in good proportion and with common-sense perspective, and (2) bringing the content, both in its detailed and its larger phases, thoroughly up to the best thought and practice available to us in 1942.

The present edition has taken advantage of constant responsible contact on the part of the authors with the technology of measurement and evaluation. The authors of the previous editions have invited another experienced author to join them in the preparation of the present edition. This edition intends to earn the confidence of active students and workers in measurement for its day which previous editions so amply earned in their day.

Classroom teachers, supervisors, and those in training for teaching or supervision will find this book a carefully-written fundamental text on the principles of measurement and evaluation in education. The main contribution of the book to the growing literature on measurement is, I think, not so much in novel points of view or advances in the technicalities of test construction as in plainness of exposition and balance in treatment of many points which to some would otherwise

seem over technical. It should be thought of as a first book in measurement and evaluation for those who at the time of studying it know very little, if anything, about measurement in education and its application to the problems of improving classroom instruction.

A careful examination of the book will reveal not only excellence of content, as such, but also an effective learning instrument for students. A major tendency of collegiate teaching is to stress factors of presentation as well as value of content. Hence a liberal supply of practice problems and exercises requiring the use of critical judgment on the part of the student constitute an essential part of this book. The instructor may sometimes deem it wise to supply occasions for the more actual "doing of the job itself" than exercises in a single volume can conveniently provide to meet this further need.

Professor Greene has revised his *Work-Book in Educational Measurements*, in collaboration with Professor John R. Crawford of the University of Maine. The present book and the *Work-Book* supply to the college teacher and his students coordinated learning units in the field of educational measurements. Not only is the student taught, but he is given sufficient carefully-graded practice in measurement that upon mastery of the books he can face the problems of measurement in actual school situations with considerable confidence in his ability to solve such problems with success.

Teachers must have an understanding of educational tests and the interpretation and use of test results. This book is definitely designed to meet in a frank and reasonable way the need for such training. It is a basic text for classroom use in courses in educational tests and measurements. It is also well adapted for use in connection with extension classes and correspondence courses.

Copies of previous editions of this book have been used for ready reference by directors of measurement, by supervisors of subject matter who have had to do their own measuring, and by building principals and classroom teachers. We expect the present edition to be an even more satisfactory ready reference book than were the previous editions.

F. B. KNIGHT

CONTENTS

CHAPTER	PAGE
I. INTRODUCTION	1
I. What Tests Are	2
II. The Meaning of Evaluation	6
III. Organization of this Book	7
II. TYPES OF EDUCATIONAL AND MENTAL TESTS	10
I. General Classification of Tests	10
II. Educational Tests	14
III. Intelligence Tests	26
IV. Personality Instruments	30
III. DEVELOPMENT OF EDUCATIONAL AND MENTAL TESTING	36
I. Measurement to 1800	37
II. Educational Tests from 1800 to 1900	38
III. Intelligence Testing from 1800 to 1900	41
IV. Intelligence Tests from 1900 to the Present	43
V. Educational Tests from 1900 to the Present	45
VI. Personality Tests from their Origins to the Present	49
IV. CRITERIA OF A GOOD EXAMINATION	52
I. Validity	52
II. Reliability	61
III. Adequacy	63
IV. Objectivity	66
V. Administrability	68
VI. Scorability	69
VII. Comparability	69
VIII. Economy	70
IX. Utility	71
V. CONSTRUCTION OF STANDARDIZED TESTS	74
I. Meaning of Standardization	74
II. Establishing Validity of Test Content	75
III. Constructing and Validating Test Items	76

CHAPTER		PAGE
	IV. Constructing Equivalent Forms . .	81
	V. Deriving Test Norms .	84
	VI. Establishing Final Validity and Reliability	93
	VII. Preparation of Final Test Materials .	95
	 VI. USING STANDARDIZED TESTS IN THE CLASSROOM	98
	I. Instructional Uses of Achievement Tests	99
	II. Planning the Testing Program . .	105
	III. Selecting the Tests	109
	IV. Administering the Tests	114
	V. Scoring the Tests	118
	VI. Analyzing the Results of Testing . .	126
	VII. Interpreting the Results of Testing	126
	 VII. USING ORAL AND ESSAY EXAMINATIONS IN THE CLASSROOM	130
	I. Classroom Testing	130
	II. The Oral Examination	131
	III. The Essay Examination	133
	IV. Improving the Essay Examination	143
	 VIII. CONSTRUCTION AND USE OF INFORMAL OBJECTIVE TESTS	151
	I. Characteristics of Classroom Testing	151
	II. Advantages and Limitations of the Informal Objective Examination	155
	III. Construction and Use of Informal Objective Tests	160
	IV. Simple Recall Items	170
	V. Completion Items	172
	VI. Alternate-Response Items	174
	VII. Multiple-Choice Items	177
	VIII. Matching Exercises	182
	IX. Constructing Informal Objective Test Items	187
	 IX. NATURE AND MEASUREMENT OF INTELLIGENCE	199
	I The Nature of Intelligence	200
	II The Measurement of Intelligence	202
	III Types of General Intelligence Tests	205

CONTENTS		xiii
CHAPTER		PAGE
	IV. Types of Specific Intelligence Tests .	215
	V. Types of Performance Tests	218
X.	USING INTELLIGENCE TESTS IN PUPIL GUIDANCE	223
	I. General Procedures for Intelligence Testing	223
	II. Derived Results of Intelligence Testing	225
	III. Distribution of Intelligence	232
	IV. Classroom Uses of General Intelligence Tests	234
	V. Classroom Uses of Specific Intelligence Tests	237
	VI. Classroom Uses of Performance Tests	238
	VII. Derived Measures Relating Intelligence and Achievement	238
XI.	USING PERSONALITY INSTRUMENTS IN PUPIL GUIDANCE	244
	I. The Nature of Personality	244
	II. Techniques of Personality Measurement	247
	III. Measurement of Attitudes	251
	IV. Measurement of Interests	254
	V. Measurement of Emotional Adjustment	257
	VI. Measurement of Total Personality	264
XII.	THE USE OF OTHER TECHNIQUES AND TOOLS IN PUPIL GUIDANCE	269
	I. Education as Adjustment	269
	II. Guidance in Adjustment	270
	III. Cumulative Pupil Records as Adjustment Tools	273
	IV. The Use of Test Results in the Adjustment of Pupils	276
	V. The Use of Other Techniques in the Adjustment of Pupils	284
XIII.	TESTS IN DIAGNOSIS AND REMEDIAL TEACHING	289
	I. The Place of Diagnosis and Analysis	289
	II. The Place of Remedial Instruction	296
XIV.	MEASUREMENT AND REMEDIATION IN ARITHMETIC	304
	I. Course Content and Organization in Arithmetic	304
	II. Measurement of General Achievement in Arithmetic	309

III. Diagnostic Testing in Arithmetic Skills	313
IV. Testing of Problem-Solving Ability	316
V. Remedial Instruction in Arithmetic	318
 XV. MEASUREMENT AND REMEDIATION IN THE RE- CEPTIVE LANGUAGE ARTS	326
I. Identification of Major Reading Abilities	327
II. General Analysis and Diagnosis of Reading Dis- abilities	331
III. Determination of Reading Readiness	333
IV. Analysis and Diagnosis in Oral Reading	336
V. Analysis and Diagnosis in Silent Reading	339
VI. Corrective Exercises in Reading	347
 XVI. MEASUREMENT AND REMEDIATION IN THE Ex- PRESSIVE LANGUAGE ARTS	355
I. Identification of Language Abilities	355
II. Measurement and Diagnosis of Language Abili- ties	360
III. Remedial Instruction in Language	368
IV. Importance of Measurement in Spelling	373
V. Construction of Spelling Tests	376
VI. Diagnosis and Remediation of Spelling Disabili- ties	379
VII. Importance of Measurement in Handwriting	384
VIII. Measurement of Handwriting Ability	388
IX. Diagnosis and Remediation of Handwriting	391
 XVII. MEASUREMENT IN THE SOCIAL STUDIES	402
I. Aims and Organization of the Social Studies	402
II. Measurable Qualities in History, Civics, and Geography	404
III. Standardized Social Studies Tests	408
IV. Informal Objective Tests in the Social Studies	415
V. Corrective Work in the Social Studies	418
 XVIII. MEASUREMENT AND REMEDIATION IN THE ELE- MENTARY SCIENCES	421
I. Scope of the Elementary Sciences	421
II. Limitations of Measurement in the Sciences	425

CONTENTS		XV
CHAPTER		PAGE
	III. Standardized Tests in Elementary Science	427
	IV. Informal Objective Testing in Elementary Science	432
	V. Diagnosis and Remedial Teaching in Elementary Science	439
XIX.	MEASUREMENT IN THE FINE ARTS	443
	I. Measurable Qualities in Music	444
	II. Measurement of Musical Talent	445
	III. Measurement and Remediation in Musical Achievement	448
	IV. Characteristics and Aims of Art Education	452
	V. Measurement of Art Abilities and Achievement	455
XX.	MEASUREMENT IN HEALTH AND PHYSICAL EDUCATION	462
	I. The Scope and Aims of Health Education	462
	II. Measurement and Evaluation in Health Education	464
	III. Prevention and Diagnosis in Health Education	467
	IV. The Objectives of Physical Education	468
	V. Measurement in Physical Education	469
	VI. Diagnosis in Physical Education	475
XXI.	MEASUREMENT OF GENERAL EDUCATIONAL ACHIEVEMENT	479
	I. General Measures of Achievement	479
	II. Types of General Achievement Batteries	482
XXII.	SUMMARIZING THE RESULTS OF TESTING	493
	I. Classifying and Tabulating Test Scores	494
	II. Measures of Central Tendency	503
	III. Measures of Variability	515
	IV. The Relationship of Test Scores	528
	V. Assignment of Relative and Percentile Ranks	541
	VI. Summary	547
XXIII.	INTERPRETING THE RESULTS OF TESTING	549
	I. Test Scores	549
	II. Types of Derived Scores	552

III.	Practical Uses of the Standard Deviation	560
IV.	Practical Uses of the Correlation Coefficient	564
V.	The Use of Norms for Interpreting Test Results	568
XXIV.	USING THE RESULTS OF TESTING	583
I.	Pupil Adjustment through Measurement	583
II.	Measurement of Progress and Improvement by Standardized Tests	588
III.	Measurement by Informal Objective Tests	589
XXV.	TESTS AND THE CLASSROOM TEACHER	598
I.	The Need for Measurement	598
II.	Objective Non-Standardized Tests	600
III.	Standardized Educational Tests	601
IV.	Interpretation of Test Results	603
V.	Practical Aspects of Classroom Measurement	604
GLOSSARY	611
INDEX	627

LIST OF TABLES

TABLE	PAGE
I. Statistical Evidence of Validity of the Powers General Science Test	59
II. Scores Assigned by Ten Teachers to an Essay and a True-False Examination over the Same Material in Civics	67
III. Discriminative Power of Test Items in Per- centage of Failure by Superior and Inferior Groups	81
IV. Age and Grade Equivalents for Part or Sub- Test Scores on the Metropolitan Arithmetic Tests	87
V. Rate Percentile Scores for Schrammel-Gray High School and College Reading Test	88
VI. Distributions of Scores on the Iowa Grammar Information Test	89
VII. Percentile Norms (Based on Data of Table VI)	90
VIII. Distribution of Intelligence in a Ninth-Grade Class in Terms of Average Grade Placement	104
IX. Shifting Standards of Expectancy	137
X. Distribution of Intelligence Quotients in a Nor- mal Population	233
XI. Record of Test Scores of Pupil A.L.C.	294
XII. Scope of Drill Units in Whole Numbers, Econ- omy Remedial Exercises	320
XIII. Analysis of Problem Solving, Economy Problem- Solving Exercises	322
XIV. Diagnostic and Remedial Chart Punctuation and Usage	370
XV. Grade Standards in Handwriting	386
XVI. Handwriting Quality and Speed Standards	387
XVII. Percentage of Mention of Various Aims of Science Teaching Found in 130 Sources (Adapted from Noll)	423
XVIII. Summary of Stanford Achievement Tests	483

TABLE		PAGE
XIX.	Summary of Cooperative Achievement Tests for the Junior High School	485
XX.	Summary of Iowa Every-Pupil Tests of Basic Skills, Elementary	486
XXI.	Summary of Unit Scales of Attainment	487
XXII.	Arrangement of Items in Unit Scales of Attainment	488
XXIII.	C-Score Equivalents for English Usage Test of Unit Scales of Attainment, Division 3	489
XXIV.	Summary of Modern School Achievement Tests	490
XXV.	Arithmetic Test Scores of 37 Seventh-Grade Pupils in Alphabetical Order of Pupils' Names	495
XXVI.	Arithmetic Test Scores of 37 Seventh-Grade Pupils in Descending Order	495
XXVII.	Arithmetic Test Scores of 37 Seventh-Grade Pupils in Frequency Distributions	496
XXVIII.	Relation of Range and Size of Class-Interval	498
XXIX.	Arithmetic Test Scores of 37 Seventh-Grade Pupils in a Grouped Frequency Distribution	500
XXX.	Computation of the Arithmetic Mean for the Grouped Frequency Distribution of 37 Arithmetic Test Scores	506
XXXI.	Computation of the Median for the Grouped Frequency Distribution of 37 Arithmetic Test Scores	511
XXXII.	Data Showing Identical Means but Unlike Variability	515
XXXIII.	Computation of the Quartile Deviation for the Grouped Frequency Distribution of 37 Arithmetic Test Scores	520
XXXIV.	Computation of the Standard Deviation from Ungrouped Data (Data for Class A from Table XXXII).	524
XXXV.	Computation of the Standard Deviation for the Grouped Frequency Distribution of 37 Arithmetic Test Scores	526
XXXVI.	Pairs of Test Scores	531

LIST OF TABLES

TABLE		XIX PAGE
XXXVII.	Illustrating the Computation of xy Products	534
XXXVIII.	Correlation Table Showing Relation of Rate and Comprehension as Measured by a Cer- tain Reading Test	535
XXXIX.	Percentage of Forecasting Accuracy for Specific Values of r	538
XL.	Relative Ranks	542
XLI.	Computation of Deciles	544
XLII.	Interpretation of Percentiles	545
XLIII.	Standard Deviation Technique for Assigning Class Marks	562
XLIV.	G-Scores for Thorndike-McCall Reading Scales	569
XLV.	Grade Norms for Intelligence and Achievement Tests	571
XLVI.	Age Equivalents of Equated Scores for Parts and Total of Stanford Achievement Test	573
XLVII.	Pupil Record from Compass Diagnostic Test No. VII	574
XLVIII.	Grade Norms for Compass Diagnostic Test No. VII	575
XLIX.	Age Equivalents of Total Scores for Compass Diagnostic Test No. VII	575
L.	Percentile Norms for High School Seniors, Rinsland-Beck English Usage Test	577
LI.	Mental Age Values Corresponding to Standard Scores, Pintner General Ability Tests	578
LII.	Scores from a Ninth-Grade Class on the Iowa Algebra Aptitude Test	587
LIII.	Marks Assigned to a Sixth-Grade Geography Paper by 557 Individuals	591
LIV.	Suggested Point Values Corresponding to Letter Marks	596

LIST OF FIGURES

FIGURE	PAGE
1. Analysis and Diagnosis	22
2. Handwriting Scale of Progressive Achievement Tests	23
3. The Principle of Sampling	65
4. Graphic Representation of the Median, 75th and 25th Percentiles Given in Table VII	91
5. Diagnostic Profile Chart for Progressive Reading Test	100
6. Individual Educational Chart for Gray-Votaw General Achievement Tests	101
7. States Having State-Wide Cooperative Testing Programs, 1933	110
8. Sample Strip Scoring Keys for Metropolitan Achievement Tests	119
9. Cutout Scoring Stencil for Iowa Every-Pupil Test of Basic Skills	120
10. Sample Test Card Marked by Krexite	121
11. International Test Scoring Machine	123
12. Samples of Machine-Scored Answer Sheets	124
13. Illustrations of Multiplex Quick-Score Grader	125
14. The Influence of Limited Sampling on Test Scores	134
15. Marks Assigned to an English Examination by 142 Teachers	135
16. The Influence of Extensive Sampling on Test Scores	156
17. Tests of the Pintner-Paterson "Long" Performance Scale	220
18. Percentage of Persons in a Normal Population at Different Levels of Intelligence	233
19. Sample of American Council on Education Cumulative Record Form for Elementary School Pupils	275
20. Sample of Diagnostic Profile Chart for Progressive Achievement Tests	277
21. Sample Graphic Record of Pupil Progress as Measured by Metropolitan Achievement Tests	279
22. Sample Profile Chart for California Test of Personality	281

LIST OF FIGURES

xxi

FIGURE

PAGE

23. Sample Class Analysis Chart for Metropolitan Achievement Tests	282
24. Mid-Points and Limits of Class-Interval	501
25. Illustrating the Principle of Moments of Forces	504
26. Illustrating the Calculation of the Mean	507
27. Illustrating the Assumptions Concerning the Distribution of Scores in Class-Intervals in the Computation of the Arithmetic Mean and Median	514
28. Illustrating the Need for Measures of Variability	516
29. Comparison of Standard Deviation (σ) and Quartile Deviation (Q)	522
30. Percentile Graph .	546
31. Pupil's Graphic Record	572

LIST OF PROBLEMS

TABULATING TEST SCORES

PROBLEM	PAGE
1. Tabulating Algebra Test Scores	502
2. Tabulating History Test Scores	502
3. Tabulating Reading Test Scores	503

COMPUTING THE ARITHMETIC MEAN

4. Computing the Arithmetic Mean	509
5. Computing the Arithmetic Mean	509
6. Computing the Arithmetic Mean	509

COMPUTING THE MID-MEASURE AND MEDIAN

7. Finding the Mid-Measure	513
8. Finding the Mid-Measure	513
9. Computing the Median	513
10. Computing the Median	513

COMPUTING THE QUARTILE DEVIATION

11. Computing the Quartile Deviation	520
12. Computing the Quartile Deviation	521

COMPUTING THE STANDARD DEVIATION

13. Computing the Standard Deviation	528
14. Computing the Standard Deviation	528

CALCULATING THE CORRELATION COEFFICIENT

15. Calculating the Pearson Product-Moment Coefficient of Correlation	540
-----------------------------------------------------------------------	-----

ASSIGNING RELATIVE RANKS

16. Assigning Relative Ranks	543
17. Assigning Relative Ranks	543

LIST OF PROBLEMS

xxiii

COMPUTING AND GRAPHING PERCENTILE DATA

PROBLEM	PAGE
18. Computing Percentiles	546
19. Constructing a Percentile or Ogive Curve	546

COMPUTING T-SCORES

20. Computing T-Scores	560
------------------------	-----

ASSIGNING MARKS FOR TEST SCORES

21. Assigning Marks for Test Scores	563
-------------------------------------	-----

ESTIMATING TEST RELIABILITY

22. Estimating Test Reliability by the "Chance-Half" Correlation Method	567
23. Estimating Test Reliability by the "Footrule" Method	567

INTERPRETING TEST SCORES

24. Finding Grade Equivalents from Test Scores	579
25. Finding Age Equivalents from Test Scores	579
26. Finding Percentiles from Test Scores	580
27. Finding Intelligence Quotients from Test Scores and Pupil Ages	580

MEASUREMENT AND EVALUATION
IN THE
ELEMENTARY SCHOOL

Measurement and Evaluation in the Elementary School

CHAPTER I

INTRODUCTION

The purpose of this chapter is to introduce the reader to the following points underlying the idea of measurement in education and with respect to the contents of this book

- a.* Measurement in education not new.
- b.* Characteristics of educational tests.
- c.* Purposes tests do and do not serve.
- d.* Importance of examinations to the school and the teacher.
- e.* Organization of this volume.

Educational Measurement Not a New Idea. Educational tests and the information resulting from their use in the classroom are coming to be almost universally identified with good teaching practice. In many ways teachers have always endeavored to measure the progress of their pupils toward an educational goal and to diagnose revealed defects in instruction. This measurement of progress and the accompanying attempts at diagnosis have been in the past largely a matter of observation on the part of the teacher. Yet the recent development of educational tests may be regarded as an extension and improvement of an old practice. However, since objective tests are more precise and exact than the ordinary teacher's examination, they thus accomplish the purposes of measurement much more accurately. They permit the setting up of specific and objective goals of achievement which are based upon the actual attainments of children under typical school conditions. They often provide answers to such specific questions as "How much should be expected of a given class or pupil under certain school

conditions?" or "How much progress should be made in a given subject in a given period of time?"

I. WHAT TESTS ARE

Measurement in Education. In most fields of human endeavor the most efficient results are attained when the worker has definite goals toward which to work and dependable instruments for determining progress. A definite aim enables the worker to direct his efforts toward the particular task to be accomplished. By the proper use of instruments for measuring results it is possible for the worker to know what he has accomplished. Thus a reliable and analytic silent reading test will give to the teacher a measure of his relative success in developing silent reading skills in his class. Accurate measuring instruments also aid in discovering when emphasis has been misplaced. For example, a pupil who, in his elementary school work, has been given an unusual emphasis on oral reading may satisfactorily pronounce words appearing on the printed page, but may be sadly lacking in ability to get meaning from these same words.

Measuring instruments also make it possible for the worker to resort to experimental methods and thus to learn definitely whether materials and methods are effective. This is as true in the field of teaching as in other fields. Without specific aims the teacher cannot plan his work effectively. He cannot know, except in an indefinite way, what he is to do. A teacher without specific aims is like a person who starts out to walk to a certain place without any idea of which direction he is to take or how far away his destination may be. If, on the other hand, the goals of instruction are clear-cut and accurate, means of determining progress are provided, and the probability of a timely arrival at the goal is greatly increased.

Characteristics of Educational Tests. Modern educational tests differ in several respects from the typical examination of the discussion type such as teachers commonly construct. First, the questions or exercises which are used in educational tests are often much more carefully selected to

coincide with the purpose for which the test is designed than is true of questions in the ordinary essay examination. For example, an objective or a standard test having for its purpose the measurement of ability to locate the states of the United States contains only such exercises as relate specifically to that purpose. The traditional essay-type examination frequently contains exercises from many fields. Second, the exercises in a carefully made educational test are commonly arranged in accordance with certain principles of test construction so as to form an accurate measuring instrument. These principles, being quite technical, are discussed in detail in Chapters V and VIII of this book, their study may be safely postponed. For the present it is enough for the reader to note that there are important principles of arrangement of items within a test which should be considered.

In the third place, the standardized test or scale is given to a large number of children of varying age and school classification, and from these results the norms are obtained. The interpretation of norms and the meaning of certain scores derived from these values are discussed in Chapter V of this book.

A fourth point of great importance, but only recently recognized and applied in connection with test development, is the fact that test scores yielded by narrow-function tests are more readily translatable into specific remedial procedures where needed. This means that a really valuable test, in addition to giving a cross-section of the situation, must diagnose, and that this diagnosis must be so specific that the results may be readily interpreted in terms of the specific kind and approximate amount of remedial attention needed.

Although the above discussion relates particularly to standardized tests, it is now recognized that the teacher can construct his own objective tests to serve purposes for which no suitable standardized tests are available or for which teacher-made tests are more satisfactory. The standard test and the teacher-made test supplement each other, and both are important in a well-rounded testing program. Certain other measurement devices of a non-test type are also of great significance in the evaluation of child behavior and of the results of classroom instruction.

What Tests Do. Probably every alert teacher has at some time earnestly desired to know whether instruction in certain school subjects was particularly superior or inferior, effective or ineffective. Many teachers have taken steps to answer this highly important question through the use of one kind of measuring instrument or another, yet relatively few classroom teachers are utilizing to the fullest extent one of the most valuable instruments at their command. Supervisors and administrators are frequently obliged to admit that they do not have adequate data on which to base their decisions as to the efficiency of a certain method of instruction, but must be guided largely by their own personal opinions. With the development of more valid and reliable measuring devices, more objective information on such questions has become available.

Standardized tests are not panaceas for all educational inadequacies, but unquestionably the scores they afford are useful in the evaluation of instruction. Test scores are objective, and may be given meaning by the process of standardization. For example, a quality score of 46 assigned to a certain handwriting sample is meaningless until it is understood that 46 points is standard quality for a fourth-grade child. If such a sample were scored as a second-grade product it might be assigned a superior mark. If scored by an eighth-grade teacher it might readily be given an inferior mark. Thus, by the use of standardized test scores, specific goals of achievement may be set up, and progress may be measured. Test norms themselves provide the basis for the more objective grade placement of pupils. Analytic and diagnostic tests make possible the discovery of pupils needing special corrective instruction. Weak spots in the course of study may be pointed out and methods of instruction may be evaluated through the critical use of educational tests and interpretation of their results.

Experience in the use of tests in many school systems leads to the conclusion that quite often there is an intangible psychological effect resulting from the administration of a series of tests in a school. The experience of the children while taking the test, and the feeling on the part of the teacher that his work is being carefully checked, are both

motivating forces making for better and more effective teaching and learning situations. This is, however, only a by-product of the use of the test and is no reason for allowing the work to stop short of a really constructive supervisory program.

What Tests Do Not Do. Standard tests are incapable in and of themselves of directly improving instruction in any subject. They merely reveal the situation. In a sense, a test might be thought of as an educational barometer. It reveals the educational-atmospheric pressure, but does not do anything about it. Perhaps the parallel is closer than at first appears. Just as low barometric readings indicate low atmospheric pressure and forecast changing weather conditions or storms, low achievement test scores may presage an unsatisfactory educational situation. As high or rising barometric records indicate fair weather, so high scores on educational tests indicate a satisfactory instructional situation.

The chief service of tests lies in their power to reveal the strengths and weaknesses of individual pupils or of the class as a whole. The use of tests must be followed by the next logical step, the development of a constructive supervisory program. It is not enough that weaknesses be revealed. They must be corrected by the use of properly constructed remedial exercises.

Significance Attached to Examinations. It requires but a casual inspection of educational practices to discover the significance which is attached to examinations by the school, as well as by the teacher, the pupil, and the parent. Pupils spend a great deal of time in preparing for and writing examinations. The school spends considerable time and money setting up an organization for the preparation and administration of examinations. Teachers devote much time to the preparation, scoring, and marking of examination papers, while parents in general set far too much store by the marks earned by their children on school examinations.

Examinations play an important part in the public relations contacts of the school. To a certain extent they carry to the parents in the community the educational purposes of the school, the aims of specific subjects and courses, and

the various emphases held important by the instructional agents of their school. Examinations in part serve as a means of revealing to both parent and pupil the basis for a pupil's scholastic rating, his promotions, failures, conditions, awards, and preparation for further educational work.

For the teacher, examinations make possible the setting up of specific objectives and provide a means of determining his efficiency in achieving them. They aid in revealing over-emphasis or wrong emphasis in teaching method and make possible the experimental evaluation of subject-matter organization.

II. THE MEANING OF EVALUATION

Several different attitudes toward the use of educational measurements in the school have held sway at various times since the objective approach to the measurement of pupil intelligence and achievement made its appearance shortly after the beginning of the twentieth century. These different attitudes or outlooks may be called by the following names: (1) testing, (2) measurement, and (3) evaluation and appraisal.

The first concept chronologically was that of testing, which considered the development of objective devices for testing intelligence and achievement of pupils to be of major importance. This attitude was doubtless the result of the early need for the development of objective instruments, as such instruments were not available in any significant quantity for some years after the concept of objectivity of tests first made its appearance in the field of education.

When objective tests became fairly numerous and classroom teachers began to use objective methods in their own examinations, attention turned more toward the use of test results and toward the development of instruments for measuring the types of instructional outcomes which do not lend themselves to ready objective measurement. This period may be characterized as one during which the major approach was that of measurement.

The quite recent development of the evaluation and appraisal concept was doubtless impelled by the increasing realization that paper-and-pencil tests can measure only a por-

tion of the outcomes of instruction and types of pupil behavior about which the teacher and other school officers need information. Therefore, the present view is that tests constitute probably the major type of evaluative instruments but that such other means of measurement as the anecdotal record, the interview, the questionnaire, the rating scale, and such tools as the individual pupil profile, the class record, the cumulative record, and the case study have a significant place in the evaluation of pupil behavior and achievement. The evaluation concept has also doubtless been stimulated by the recent attention of educators and psychologists to the whole child and his behavior. This tendency to consider the child as a whole, rather than as an individual whose behavior and abilities can be catalogued into a number of different compartments, places a responsibility upon the user of tests and other instruments of evaluation for considering the child in this broad sense. It is through the application of the evaluation concept rather than of the narrower concepts of measurement and testing that this result is most effectively obtained.

The approach may be summarized by the statement that teachers participating in summer workshops where evaluation was under consideration were "using instruments of evaluation to discover not whether pupils had done their work but what the work had done to the pupils."¹

III. ORGANIZATION OF THIS BOOK

Purpose of This Book. The purpose of this book is two-fold: (1) to interest the student of elementary education in the possibilities of measurement and evaluation in education, and (2) to stimulate the elementary school teacher and supervisor to make more effective use of tests and other evaluative devices as integral parts of enlightened teaching practice. To accomplish this two-fold purpose the reader is gradually introduced to the meaning and possibilities of measurement through the examination of some of the well-known current classroom practices. Tests are classified into

¹ Kenneth L. Heaton, William G. Camp, and Paul B. Diederich, *Professional Education for Experienced Teachers*, p. 120. University of Chicago Press, Chicago, 1940.

their major types in Chapter II and brief descriptions are given of each type. Chapter III briefly outlines the important steps in the development of educational and mental testing.

Chapters V to XII present the methods of constructing and the values and uses of the major types of tests and evaluative techniques to the teacher—standardized tests in Chapters V and VI, teacher-made tests in Chapters VII and VIII, intelligence tests in Chapters IX and X, personality instruments in Chapter XI, and broad evaluative techniques of a non-test nature in Chapter XII.

The teacher successful in the use of tests in his teaching must be able to secure an accurate summary and interpretation of results. For the purpose of developing these skills, Chapters XXII and XXIII are presented. The simple statistical procedures, such as tabulating test scores and calculating medians, arithmetic means, quartiles, and the like, will be found especially helpful in interpreting results where fairly large numbers of cases are concerned. The discussion of derived scores in Chapter X should also be useful in making an analysis and interpretation of test results. Experience with practical problems of this nature is provided by means of a series of exercises which are scattered through Chapters XXII and XXIII.

Chapter IV, which discusses at some length the characteristics or criteria of a good examination, is exceedingly important. It can most advantageously be studied after Chapters XXII and XXIII have been taken up, for a comprehensive understanding of the two most important criteria of a good examination depends upon the ability to interpret correlation coefficients. However, a careful reading of pages 529 to 531 and pages 537 to 539 in Chapter XXII and a brief survey of pages 564 to 567 in Chapter XXIII will perhaps sufficiently acquaint the student with the meaning of correlation for the immediate purposes of Chapter IV.

Those especially interested in following to its logical conclusion the use of tests in the classroom will wish to study with particular care Chapters XIII to XXI and XXIV and XXV, in which are presented the possibilities and the practical methods of using test results for diagnosing the learning

difficulties of pupils, and the inauguration of preventive and remedial instruction in important elementary school subjects.

Study Aids. The student who is genuinely interested in improving his understanding of many of the points presented in this volume will find much profit in the careful preparation of the discussion exercises at the end of each chapter. Those who are still more deeply interested in, and wish to pursue further, the problems of measurement in education will find the selected references at the close of each chapter of particular value. Because the field of educational measurements is so rich and the essential material is so extensive, it is impossible to compress into the few pages allotted to this book even much of that which is considered by many to be fundamental.

Teachers themselves expert in the technique of learning know that passive reading, while yielding information and appreciation, does not develop easy, dependable skill in doing the thing described. To provide the opportunity for the student and the teacher actually to secure a more complete mastery of certain of these techniques, a *Work-Book in Educational Measurements* has been prepared as a companion volume for this treatment. In this *Work-Book* the reader may solve actual and practical problems of the type which the classroom teacher and supervisor face. Mastery of this text and a careful working of the sampling of projects in the *Work-Book* will practically guarantee to the reader an actual concrete experience with the major problems of a dynamic testing program calculated to be of the greatest service in the improvement of typical classroom situations.

TOPICS FOR DISCUSSION

1. What specific evidence do you find that the idea of measurement in education is not entirely new but has been in the minds of teachers for many years?
2. How far, in your opinion, is the classroom teacher responsible for the understanding and use of educational tests?
3. Indicate several major characteristics of educational tests.
4. Briefly distinguish between standardized tests and teacher-made tests.
5. Specify several things which educational tests, when properly used, do for the classroom teacher and his pupil.
6. Why are educational tests not panaceas for all weaknesses of the school?

CHAPTER II

TYPES OF EDUCATIONAL AND MENTAL TESTS

This chapter presents a classification of educational and mental tests and discusses the major characteristics of each type of test.

- a.* General classification of educational and mental tests.
- b.* Types of educational tests.
- c.* Standardized achievement tests.
- d.* Teacher-made or classroom tests.
- e.* General intelligence tests.
- f.* Specific intelligence or aptitude tests.
- g.* Performance tests.
- h.* Personality instruments.

The measurement and evaluation of the total personality, ability, and achievement of pupils involve the use of a wide variety of tests and other devices which cannot properly be called tests. The types of measuring instruments known as tests are discussed in this chapter and in many chapters later in this volume, while the other types of measuring instruments are treated primarily in Chapters XI and XII.

I. GENERAL CLASSIFICATION OF TESTS

Educational and Mental Tests. Modern tests are so varied in type that it is extremely difficult to classify them clearly. Tests can be classified in terms of their forms, their origins, their functions, and their content. In the three major sections of this chapter, tests are first classified broadly by function—educational, intelligence, and personality—and within major divisions are classified by whatever pattern seems most likely to familiarize the student with their major characteristics.

Educational tests have as their primary function the measurement of the results or effects of instruction and learning. On the other hand, *intelligence tests*, or psychological examinations, have as their purpose the measurement of pupil intelligence or mental ability in a large degree without refer-

ence to what the pupil has learned either in or out of school. *Personality tests* attempt to measure such intangible aspects of behavior as attitudes, interests, and emotional adjustment.

There is not complete uniformity of terminology with respect to educational tests and mental tests. Although the former have a commonly-accepted meaning, the latter are thought variously to include educational, intelligence, and personality tests, to include intelligence and personality but not educational tests, and even to mean the same thing as intelligence tests. The practice of Freeman¹ and others is to distinguish between educational and mental tests and to consider the latter as including intelligence and personality tests. As this distinction appears to be most satisfactory for the purposes of this book, it will be followed here and throughout the volume.

Tests, Scales, and Scaled Tests. Objective tests can be classified in a manner which cuts across the three fields of educational, mental, and personality testing—into tests and scales, and also scaled tests. This distinction is of some value, but at times it results in confusion since certain types of objective tests resemble scales or contain certain features of scales as an essential part of their construction.

In general terms, a *test* is an instrument designed for the measurement and evaluation of any knowledge, quality, or ability. It may measure degree or amount of achievement, mental abilities, or even such indefinite qualities as personality and character traits. It may be made up of items of similar difficulty, or it may be composed of items of uniformly increasing difficulty or value. Ordinarily the test is used in the classroom by the pupils.

The term *scale* is used to designate a series of objective forms of exercises or definite samples or products of different quality which, by means of a rather technical statistical procedure, have been arranged in a definite order or position, usually in ascending order of difficulty or merit. In a scale, the difference in value or difficulty or quality which separates each exercise from the one just below it on the scale is equal to the difference between it and the exercise next above it on

¹ Frank N. Freeman, *Mental Tests Their History, Principles, and Applications* (Revised Edition), pp 13-22. Houghton Mifflin Co., Boston, 1939.

the scale. That is, the exercises are equally spaced on a scale of value, of difficulty, or of quality. Usually the scale is employed by the teacher as an aid in the evaluation of the particular product.

A *scaled test* combines certain properties of the test and the scale. If the items in a test are arranged in order of increasing difficulty, the instrument is a scaled test. The process of determining the difficulty of test items and arranging them in an ascending order on that characteristic is called *scaling*.

The scaled test is illustrated by a sampling of exercises taken from the elementary science section of the *Unit Scales of Attainment*. These exercises represent the first five and the last five items in the part designed for use in Grades 7 and 8.

One of the inconsistencies of testing terminology is that all three types of instruments mentioned above—tests, scales, and scaled tests—are commonly included in general usage under the term “tests.”

Rate and Power Tests. *Rate Tests.* A rate test usually consists of exercises approximately equal in difficulty. Ordinarily such a test contains so many items that no pupil is able to finish in the working time allowed. The number of exercises of this uniform difficulty finished correctly in the specified time is taken as the pupil's rate of work. Thus, rate tests are measures of the speed and accuracy at a given level of difficulty with which a pupil is able to respond to certain standardized exercises of a uniform nature. Usually the exercises are relatively easy and there is little or no question about the pupil's ability to do them. The quality of his performance may be expressed in terms of the percent of the exercises done correctly.

Power Tests. The power tests, or so-called scaled tests, measure a pupil's ability to do more and more difficult exercises within a given field of subject matter. In such tests, achievement is expressed in terms of the difficulty of the exercise or activity which the pupil is just able to perform, although few modern tests are scored on this basis. A power test consists of a series of exercises arranged in ascending order of difficulty. Usually no measure of the pupil's rate

EXCERPTS FROM UNIT SCALES OF ATTAINMENT, ELEMENTARY SCIENCE²

ELEMENTARY SCIENCE

Directions. Read these two sentences carefully

A. The sun rises in the 1 evening 2 west 3 south 4 morning 5 north A. 4

B. Wood comes from 1 lakes 2 trees 3 mines 4 river beds 5 plants B.

You see that there are five possible answers in each sentence. Only one answer is right. In the first sentence the right answer is morning, so a line is drawn under morning, and the number in front of it, 4, is put at the end of the line.

Now look at the second sentence above and listen to the next directions.

In each of the following sentences you are to find the right answer, draw a line under it and then put the number that is in front of it at the end of the line, just as in the samples above

- 1 Enamel is a part of one's 1 heart 2 brain 3 lungs 4 teeth 5 intestines 1
- 2 A food that contains much starch is 1 potatoes 2 lettuce 3 tomatoes 4 spinach 5 cabbage 2
- 3 The part of an electric circuit that burns out when there is too much current is the 1 faucet 2 valve 3 damper 4 switch 5 fuse 3
- 4 A bird that usually builds its nest on the ground or in low bushes is the 1 robin 2 swallow 3 meadow lark 4 wren 5 bluebird 4
- 5 The part of a furnace pipe which controls the draft is a 1 faucet 2 damper 3 valve 4 switch 5 fuse 5
86. The saliva 1 dissolves fats 2 dissolves proteins 3 changes fats to sugar 4 changes starches to sugar 5 changes proteins to fats 36
87. Leaves with five points or lobes are found on the 1 oak 2 maple 3 elm 4 ash 5 willow 37
88. The light fleecy clouds are called 1 cumulus 2 stratus 3 gnomon 4 nimbus 5 cirrus 38
89. The amount of work done is measured by 1 time taken 2 force and time 3 force and distance 4 distance and time 5 force, time and distance 39
40. Bile is secreted by the 1 liver 2 kidneys 3 appendix 4 stomach 5 intestines 40

Number right. ..

End of Elementary Science test Look over your work.

of work is secured, because the time allowed in such a test is more than enough for all subjects to complete as many of the exercises as they are able to do. In actual practice, however, the factors of rate and quality are combined by taking as the pupil's score the number of exercises of increasing difficulty (or value) done correctly in the specified time. Theoretically, a pupil's score on a power test is the degree of difficulty of the most difficult exercises he is able to do with a specified degree of accuracy. Such a score is laborious to

² August Dvorak and M. J. Van Wagenen, *Unit Scales of Attainment, Elementary Science*, Division 3. Published by Educational Test Bureau, 1933

compute from a pupil's performance, and for this reason the number of exercises done correctly is generally taken as his score.

If *rate tests* may be compared to running a race in which a series of hurdles of uniform height are to be cleared, *power tests* may be compared to a race in which the hurdles are very low at the start but become gradually higher and higher as the race goes on until no one is able to clear them. In the first case, the score (rate) would be expressed in terms of the number of hurdles cleared in a specified time. In the second case, the score (power) would be expressed as the height of the highest hurdle which the individual was just able to clear.

Most of the more recent tests are really hybrids resulting from the combination of the power idea and the rate idea in testing. That is, they are made up of scaled items arranged in ascending order of difficulty, but the resulting scores representing accomplishment are expressed in terms of the number of items responded to correctly in the specified working time regardless of difficulty.

II. EDUCATIONAL TESTS

Considered as educational tests here and throughout this book are all tests designed to measure what the individual has learned both in and out of the school. It is obviously impossible to be certain concerning the exact proportions of the attainments of a school pupil which are the result of direct classroom instruction, of the by-products of classroom and other school activities, and of the wide range of his out-of-school experiences. Tests which are designated for measuring the outcomes of instruction are intended to test primarily classroom learning. However, a rather wide variety of tests for types of abilities not definitely taught in any one classroom or even in the school should be considered educational tests, for the education of the child is not confined entirely to the hours he spends in school. This broad conception of educational tests underlies the point of view presented in this volume. Various aspects of educational testing are dealt with in greater detail in Chapters V to

VIII and in the section of this volume devoted to measurement and evaluation in various subject fields.

When test devices are classified in terms of their form or structure, three types may be distinguished—(1) *Oral Questioning*, the (2) *Traditional* or *Essay Examination*, and (3) *Objective Examinations*. If examinations are considered mainly from the standpoint of the function for which they are most widely used—measurement of pupil achievement—the first of these types can be almost entirely disregarded, for *oral questioning* is at best an inefficient measurement tool for use in the classroom. This is not to be interpreted as meaning that oral questioning has no place in the school, but rather as implying that it should be used primarily as a teaching device for immediate recall.

In the *traditional* or *essay examination*, a limited number of questions (usually five to ten) are stated by the teacher as a basis for discussion by the pupil. Typically, the questions are selected by the teacher without too critical attention to the subject matter he intends them to cover, with the result that often the questions represent only a brief and incomplete sampling of the material for which the pupils should be held responsible. Furthermore, such irrelevant factors as English, including vocabulary, sentence structure, spelling, paragraphing, and composition; neatness of the paper; legibility of the handwriting; and the teacher's attitude toward the pupil operate in unknown combinations to influence the final mark. Although the essay examination has doubtless resulted in much injustice to pupils in many classrooms of the past and even of the present, it can be so used as to occupy a significant place in the measurement program of the schools.

The two *objective types* of examinations—*standardized* (*standard*) and *informal objective* (*new-type* or *objective*)—differ not at all in form. The primary difference lies in their origins and the degree of refinement to which they have been carried. Standardized tests are the work of subject matter and test specialists, are intended for wide use, and are accompanied by norms, whereas informal objective tests usually are constructed by the classroom teacher, or at least within the school system, and are for local use only.

Both of these examination types are marked by two important features: (1) brevity of pupil response, and (2) absence of personal judgment in the scoring of the examinations. The pupil indicates his response by such simple physical reactions as underlining a word, encircling a number, filling an answer space, or writing a word or short phrase in an indicated place. Because of this brevity of response, a wide range of material can be sampled in a short space of time, and the results can be scored both quickly and accurately by persons who may even be unfamiliar with the material tested.

Standardized Achievement Tests. A test is standardized when (1) it is composed of exercises which have been selected in the light of current teaching emphasis and curricular content, when (2) these exercises have been statistically evaluated as to innate difficulty, and when (3) the test itself is accompanied by norms permitting the interpreting of the results of pupil reactions to the test in terms of levels of accomplishment. Standard achievement tests are of value in making comparisons of a class with general norms and in comparing groups in different local schools with one another and with groups from other school systems.

Survey and Prognostic Tests. These two types of tests serve very different purposes and are constructed on very different lines, but they may both be considered general tests in the sense that their functions demand resulting scores which have general significance rather than highly specific or analytical meaning.

Survey tests are instruments which measure general achievement in certain subjects or fields of knowledge. They test skills and abilities which are relatively independent of one another. Thus, a survey test might measure achievement in first-year algebra. Another, and broader, survey test might measure ability in all areas of mathematics at the high school level. A still broader survey test might measure abilities in all of the major areas of the secondary school course of study.

The characteristics of survey tests designed for use in a subject-matter field are shown in the accompanying arithmetic examples reproduced from the *Iowa Every-Pupil Test of Basic Arithmetic Skills* for Grades 6 to 8. These

SAMPLE SECTION FROM IOWA EVERY-PUPIL TESTS
OF BASIC SKILLS IN ARITHMETIC³

PART II SECTION A WHOLE NUMBERS AND FRACTIONS

Directions Do your work right on this page. Copy your answers in the boxes provided on the answer sheet. If there is a remainder in a division example, be sure to write that remainder in the box with the answer. Fractions should be written in simplest form.

41 Add $\begin{array}{r} 736 \\ 618 \\ 422 \\ 907 \end{array}$	47 Multiply $\begin{array}{r} 3070 \\ \times 208 \\ \hline \end{array}$	52 Divide $6 \div \frac{2}{3} =$	58 Subtract $\begin{array}{r} 14\frac{3}{4} \\ - 7\frac{1}{10} \\ \hline \end{array}$
42 Subtract $\begin{array}{r} 4001 \\ - 3392 \\ \hline \end{array}$	48 Divide $87 \overline{)4265}$	53 Add $5\frac{1}{3} + 3\frac{4}{5} =$	59 Multiply $2\frac{1}{3} \times \frac{5}{8} \times 4\frac{2}{5} =$
43 Multiply $\begin{array}{r} 689 \\ \times 900 \\ \hline \end{array}$		54 Subtract $\begin{array}{r} 7\frac{1}{3} \\ - 6\frac{2}{3} \\ \hline \end{array}$	
44 Divide $73 \overline{)1491}$	49 Add $\frac{5}{8} + \frac{5}{6} =$	55 Multiply $15 \times 3\frac{2}{3} =$	61 Subtract $\begin{array}{r} 9\frac{2}{6} \\ - \frac{5}{9} \\ \hline \end{array}$
45 Add $\begin{array}{r} 38 \\ 16 \\ 80 \\ 40 \\ 58 \\ 82 \\ 9 \\ 6 \\ 48 \\ 44 \end{array}$	50 Subtract $\begin{array}{r} 32 \\ - 5\frac{7}{8} \\ \hline \end{array}$	56 Divide $\frac{3}{8} \div \frac{1}{4} =$	62 Subtract $\begin{array}{r} 15 \\ - 12\frac{3}{4} \\ \hline \end{array}$
46 Subtract $\begin{array}{r} 3811 \\ - 2045 \\ \hline \end{array}$	51 Multiply $\frac{3}{5} \times \frac{1}{4} =$	57 Add $\begin{array}{r} 2\frac{7}{8} \\ \frac{3}{5} \\ 4\frac{3}{8} \end{array}$	63 Divide $5\frac{2}{3} \div 8\frac{3}{8} =$

(Do not turn to the next page until you are told to do so.)

tests cover a wide variety of arithmetic knowledges and skills.

Prognostic tests are intended for use in the prognosis or prediction of future success in specific subjects of the school curriculum. As they usually test the background skills and

³ H. F. Spitzer, *Iowa Every-Pupil Tests of Basic Skills, Test D, Basic Arithmetic Skills*, Advanced. Published by Houghton Mifflin Co., 1940.

abilities found to be prerequisite for success in the particular subject, prognostic tests are most common among subjects in which success can be rather well defined in terms of certain basic abilities. They also frequently test some of the aptitude factors which are not directly dependent upon previous training of a specific type. Therefore, prognostic tests, probably most closely related to aptitude tests, are more properly classified as educational tests than as intelligence tests, although they unquestionably do measure certain aspects of intelligence.

Diagnostic and Analytic Tests. Tests of these diagnostic and analytic types are intended for the separate measurement of rather specific aspects of achievement in a single subject or field. Diagnostic tests measure somewhat narrower aspects of achievement than do analytic tests, so they may be thought of as serving specific and general diagnostic functions respectively.

Diagnostic tests yield measures of highly related abilities underlying achievement in a subject. They are designed to identify particular strengths and weaknesses on the part of the individual child, and within reasonable limits to reveal the underlying causes.

The relation of each skill to other skills and to the total process in the case of addition of whole numbers is shown clearly in the accompanying reproduction of Test 1 of the *Compass Diagnostic Tests in Arithmetic*. The diagnostic procedure here is based on the assumption that mastery of the total process can be no stronger than the weakest link in the chain of related skills. Accordingly, each skill called into play in the total process so far as possible is isolated and measured. For instance, each of the basic facts called for in the examples in Part 4, *Carrying in Column Addition* appears as such in Part 1, *Basic Addition Facts*. Each higher decade addition fact which appears in the total process as expressed in Part 4 also appears as such in Part 2, *Higher Decade Addition*. The column addition difficulties which appear in Part 4 also appear in Part 3 of the diagnostic test. The net result is that each of the related skills is isolated one at a time until the underlying causes of the pupil's failure to do addition of this type are revealed.

Analytic tests may be considered as general diagnostic tests. The term "diagnostic" as applied to educational tests has resulted in many misconceptions. Fundamentally, all tests may be considered diagnostic in the sense that they actually yield useful information about pupil achievement. However, the diagnosis afforded by many present-day tests is extremely general. Many so-called diagnostic tests are not diagnostic, but are merely analytic tests. This is particularly true of most of the so-called diagnostic tests in such subjects as language, reading, history, science, and the social studies.

In contrast with the specific diagnosis which appears to be possible in the case of arithmetic is the general type of analysis which seems to climax the best efforts of test makers in the fields of language, reading, and certain other highly complex subjects. Attempts to analyze language and reading with a view to the construction of diagnostic instruments immediately encounter the impossibility of relating to each other in any causal way the several phases of the subject upon which achievement in it seems to depend. For example, causal relationships have not been established among such common factors in silent reading ability as word meaning, rate of reading, comprehension of facts, ability to get the main idea, etc., so in many subjects measures of different abilities are necessarily treated as independent and unrelated aspects of total ability.

An illustration may serve to summarize the essential features of diagnostic and analytic tests. On certain points along the rim of the Grand Canyon there are look-out stations equipped with telescopes, each pointed and focused upon a specific spot of beauty or grandeur. From each of these a separate view of the beauty spots of the canyon is secured. From the composite of all of these views, a much more accurate appreciation of the total panorama is obtained, yet each view is quite independent of every other one. This is typical of the way in which tests of the analytic type operate. In distinct contrast with this example, the best way to illustrate the operation of the diagnostic type of test is to liken it to an inverted pyramid made of bricks. The removal or the crumbling of a single brick at any point in the

SPECIMEN OF COMPASS DIAGNOSTIC TESTS IN ARITHMETIC⁴



Standard Mathematical Service
COMPASS DIAGNOSTIC TESTS IN ARITHMETIC
RUCR—KNIGHT—CHENOWETH—STODOLSKER
EDITED BY G. W. MYERS

**TEST I ADDITION OF WHOLE NUMBERS: FORM A**

Name _____ Grade _____ Boy or girl? _____

Age _____ **When is your next birthday?** _____ **How old will you be then?** _____

School _____ Date _____
(Name) (City) (State)

SUMMARY OF PUPIL'S SCORE	PART 1	PART 2	PART 3	PART 4	PART 5	TOTAL
SCORES ON PARTS OF TEST						
EDUCATIONAL AGE EQUIVALENT						
GRADE EQUIVALENT OF SCORE						

PART I—BASIC ADDITION FACTS

Auld

0+3=	4+1=	0	9	5	2	3	1	7	9	8	5	7	6	4	2
1+7=	5+3=	2	1	0	3	4	1	5	2	4	4	4	2	1	6
2+2=	1+0=														
7+6=	6+4=														
2+5=	3+4=	9	4	8	0	4	2	6	5	2	4	3	7	8	8
2+1=	2+7=	5	0	6	7	2	5	6	3	4	5	7	3	3	8
0+9=	1+6=														
9+2=	3+3=														
3+9=	7+4=	9	0	1	9	2	7	1	7	1	7	5	8	3	6
5+8=	0+0=	6	9	7	7	8	6	8	8	6	7	9	0	9	3
9+1=	9+9=														
8+5=	7+5=														
2+8=	9+4=	8+3=	7+2=	Score on Part I = Number right = _____ [Total possible score = 70 points]											

PART 2—HIGHER DECADE ADDITION

Add

3	21	6	8	18	4	9	7	24	10	6	16	14	0	4	24	21	4
20	2	22	13	6	12	29	25	7	2	12	6	7	10	42	2	4	18
32	1	4	28	4	38	7	1	11	5	15	0	1	38	40	2	1	5
6	18	16	7	20	3	13	24	9	10	9	12	11	6	2	25	14	28
31	6	9	2	24	10	8	9	32	10	3	7	0	2	6	7	19	9
0	11	20	36	9	4	16	27	2	7	18	26	31	26	13	13	2	23

⁴ G M Ruch, F B Knight, H A Greene, and J W Studebaker, *Compass Diagnostic Tests in Arithmetic, Test I, Addition of Whole Numbers*. Published by Scott, Foresman and Co, 1925

Part 2—Continued

Add:

10+7=	14+2=	10+3=	20+7=	23+7=	0+20=
12+2=	11+7=	18+8=	26+6=	18+7=	22+0=

Score on Part 2 = Number right $\times 5 =$ _____
 [Total possible score = 60 points]

Part 3—COLUMN ADDITION

Add:

4	7	6	9	6	9	9	4	7	8	2	6
7	4	6	4	4	7	2	2	6	0	4	3
8	9	9	6	0	8	4	9	7	1	3	5
—	1	3	4	7	9	7	7	4	2	8	7
		4	9	5	1	5	8	8	4	1	0
		—	—	0	6	9	8	7	1	8	4
				—	—	—	3	6	2	6	7

Score on Part 3 = Number right $\times 5 =$ _____
 [Total possible score = 60 points]

Part 4—CARRYING IN COLUMN ADDITION

Add:

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
29	98	17	47	117	766	27	9132	7027	162	653		
47	13	13	4	960	982	96	5112	249	4914	5156		
—	—	—	—	—	—	—	—	—	—	—		
		67	13	517	932	77	2343	7191	9737	7380		
		49		448	84	7417	9235	67	2661			
		29		—	98	—	9059	9280	4099			
		—			17	—	427	9730				
					26	—	—	—	9223			

12. Copy and add $62+604+827+797+997=?$ (Put your work under 12 above.)
 13. Copy and add 76, 64, 00, 97, 35, and 20 (Put your work under 13 above.)

Score on Part 4 = Number right $\times 10 =$ _____
 [Total possible score = 120 points]

Part 5—CHECKING ANSWERS IN ADDITION

Directions: Some of the printed answers below are right, and some are wrong. Check each sum by adding downwards. Write your check answer on the line at the top of the example. The first one is already done correctly.

797							
213	165	6506	9270	7774	887	162	663
128	923	3157	9984	9447	756	4914	5156
456	928	4918	4763	7267	798	9737	7360
787	466	7194	7163	2848	71	67	2661
	2484	77	9650	9192	700	9280	4099
		21792	3063	30528	206	427	9730
					3618	24587	9223

Score on Part 5 = Number right $\times 10 =$ _____
 [Total possible score = 70 points]

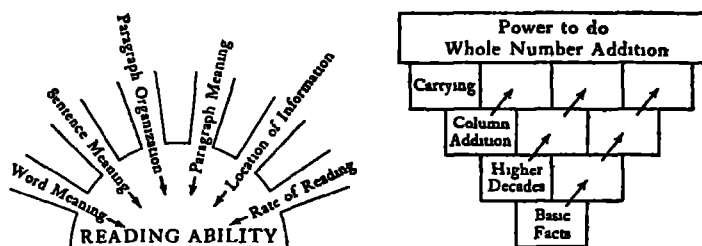


FIGURE 1. ANALYSIS AND DIAGNOSIS

wall will cause it to fall. The accompanying diagram may be helpful in clarifying the essential differences in these two types of tests.

Quality and Product Scales. Scales represent the second major division or type of standardized measures of achievement.

Achievement may be expressed in terms of ratings on *quality scales*. In such subject-matter fields as handwriting, composition, free-hand drawing, industrial shop, and agriculture, the measuring instruments for certain types of outcomes must obviously be of a different type. In these subjects the pupil produces a performance which cannot be considered as either right or wrong. The problem becomes one of describing its degree of quality. For this purpose quality scales have been constructed which consist of a series of specimen performances of the type to be described. These exhibit varying degrees of quality from the lowest to the highest. The specimens are arranged systematically in order of increasing quality. Usually the quality of each is described numerically. A quality scale of this type is used by matching the performance to be described with the specimen of the scale which most nearly resembles it in quality. Quality scales are most commonly used for the measurement of handwriting and composition abilities.

An example of a quality scale for handwriting is given in the illustration from the *Progressive Achievement Tests* for Grades 7 to 10. As is usual for quality scales, the samples are selected so that they are equally spaced along a scale of handwriting merit.

Form A and C—HANDWRITING SCALE—Form B and D

grocery doubt concert	Score 4	motion arrive believe
grocery doubt concert	6	motion arrive believe
grocery doubt concert	8	motion arrive believe
grocery doubt concert	10	motion arrive believe
grocery doubt concert	12	motion arrive believe
grocery doubt concert	14	motion arrive believe
	Score	

FIGURE 2. HANDWRITING SCALE OF PROGRESSIVE ACHIEVEMENT TESTS⁵

⁵ Manual of Directions Progressive Achievement Tests—Intermediate Battery,
1937 Revised Norms, p 9 California Test Bureau, Los Angeles

In *product scales*, limited almost entirely to spelling, all exercises of a given difficulty or value are grouped together and are properly located along a scale of difficulty. These materials are used in the construction of tests which are made up of items of known levels of difficulty for the grade in which the test is to be used, but the scales are not used in their entirety in the classroom as an actual test.

The following excerpt from the *Iowa Spelling Scales* for the eighth grade shows the words in Steps 12 and 15, for which the average percentages of misspellings are 58 and 34 respectively. On the average, the word "client" in Step 12 is misspelled by 62 percent and the word "canvass" in Step 15 by 31 percent of eighth-grade pupils.

EXCERPTS FROM IOWA SPELLING SCALES⁶

Step 12 — 58%	58%	55%
62%	all right	anticipating
client	alumni	circuit
convenient	anticipate	disappoint
council	assessment	equipped
immense	candidacy	immediately
permanent	continuous	
principle	fundamental	Step 15 — 34%
	geometry	37%
61%	girlie	definitely
accredited	physician	fraternally
characteristic	possess	
courteous	thorough	35%
losing		anniversary
scientific	57%	
	thoroughly	34%
60%		zephyr
analysis	56%	
correspondence	accompanying	33%
mortgage	acquaintance	chautauqua
Sabbath	auntie	X-ray
	originally	
59%	recommendation	31%
enthusiastic		canvass
lieutenant		
unusually		

⁶ E. J. Ashbaugh, *Iowa Spelling Scales*, Grade VIII. Published by Public School Publishing Co., 1922

Teacher-Made or Classroom Tests. The two types of educational tests which are commonly constructed by the teacher for use with his own classes are the essay and informal objective tests. Informal objective examinations are sometimes cooperatively prepared by two or more teachers for use with their several classes in the same subject, or even by several persons for use throughout a large school system. Such tests may even be printed. They are, however, informal objective examinations unless the procedures for standardization discussed in Chapter V are carried out and the tests are made available for general use to interested persons outside of the school situation where they originated.

Essay Examinations. This type of examination frequently poses a question of the *who, when, where, what, or why* type, although it may ask pupils to name, to locate, to discuss, to evaluate, to distinguish between, to define or describe, to illustrate or explain, to give reasons for or causes of, or otherwise respond to more-or-less definite issues. Too frequently such questions are so broad and involve such complexities that pupils cannot give adequate responses in the time allowed, e.g., "Discuss the causes of the First World War." Someone who apparently did not favor the essay examination facetiously suggested: "Describe the universe and give two examples."

The following illustrations are perhaps typical of essay questions which have been widely used in the past.

1. What are the major industries of New England?
2. Define a predicate adjective and illustrate its use

As all teachers are familiar with the essay examination, these illustrations of question types are sufficient at this point. A complete discussion of the essay examination and of means for improving its accuracy as a measuring instrument is given in Chapter VII.

Informal Objective Examinations. Informal objective test items are similar in form to items used in standardized tests. A tremendous variety of item types has been developed, and new adaptations are quite common. However, all objective items may be classified either as the *Recognition* or

the *Recall* type. Recognition types, of which the alternate-response, multiple-choice, and matching forms are the most common, make only indirect demands upon the initiative of the pupil, inasmuch as the factual material basic to the issue in question is stated (or misstated) in the item. Recall types, however, of which the simple recall and completion forms are probably the most common, place demands upon the initiative and frequently the memory of the pupil by expecting him to supply and state the correct answer.

The illustrations below show how a factual knowledge can be measured by the three of the above types which are most brief in form. The first two are recognition and the third is recall in form.

1. The President of the United States in 1863 was Abraham Lincoln Ⓓ F
2. The President of the United States in 1863 was (a) Ulysses S. Grant, (b) Millard Fillmore, (c) Abraham Lincoln, (d) Andrew Johnson, (e) Zachary Taylor (c)
3. The President of the United States in 1863 was (Abraham Lincoln)

The tremendous variety of informal objective examination item types and the complexity of some of them makes impossible the presentation of more than a few of the most common forms here. A comprehensive treatment of this important type of examination is given in Chapter VIII.

III. INTELLIGENCE TESTS

Intelligence tests measure what is perhaps most simply and most commonly described as ability to learn or ability to adapt oneself to new situations. Whereas achievement tests measure skills or abilities more or less directly, intelligence tests face the problem of measuring mental qualities indirectly in terms of the manner in which an individual's intelligence affects or conditions his behavior. It is sufficient here to comment upon this important distinction. Chapter IX presents more fully the problems and techniques of intelligence testing.

General Intelligence Tests. The most widely known tests of mental ability are usually referred to as general intelligence tests, although such other terms as general mental

ability tests and psychological examinations have almost identical meanings. Other terms having similar meanings are general ability tests and aptitude examinations. General intelligence tests attempt to measure mental ability broadly enough, by the use of a wide variety of test situations in scaled order of difficulty, to obtain a measure representative of the individual's mental efficiency in general.

Results from general intelligence tests have so many uses, as in educational guidance, vocational guidance, sectioning of classes, discipline, and diagnosis, that it is impossible at this point to do more than mention this fact. The uses of the results from general intelligence tests are discussed in detail in Chapter X.

Individual Intelligence Scales. Intelligence tests which can be administered to only one person at a time are known as individual intelligence examinations. Such tests require the full attention of a trained examiner. Although the techniques for administering these tests are highly standardized, the examiner modifies the procedure in various ways according to the age, ability, and even sex of the pupil being tested. These instruments are usually in scaled form, and are frequently devised to cover a wide age range, so they are often called age scales.

Individual intelligence tests only on occasional sections require the use of pencil and paper by the subject under examination. Some parts are even of a performance test nature. Many of the pupils' responses are given orally and are recorded by the examiner.

Group Intelligence Tests. Group tests of intelligence or general mental ability are usually paper-and-pencil tests which can be administered to a large group of persons at the same time. Group intelligence tests of the "omnibus" variety are ordinarily not divided into parts but have the items in mixed order with respect to the nature of the abilities they test and also sometimes with respect to their objective form. More commonly, however, group tests of intelligence have a number of different parts, each of which deals with a certain broad type of performance. Usually, but not always, both of these types of tests are given with rather rigid timing.

The accompanying sample items from the *Pintner General Ability Tests, Verbal Series*, for the intermediate grades, illustrate one of the techniques used in group intelligence tests.

SAMPLE ITEMS FROM PINTNER GENERAL ABILITY TESTS, VERBAL SERIES⁷

TEST 2 LOGICAL SELECTION						Page Verbal Intermediate 4				
<p>Directions Look at the sample that follows</p> <p>Sample A table always has — 1 flowers 2 tablecloth 3 legs 4 varnished top 5 vers .</p> <p>A table always has legs which is number 3 so the third answer space is marked in the margin</p> <p>Read each statement Find the thing it is most likely to have. Then mark the answer space in the margin which is numbered the same</p>										
1 A forest <i>always</i> has —	1 snow	2 trees	3 beasts	4 a forester	5 hunters	1	2	3	4	5
2 A sled —	1 boys	2 runners	3 ice	4 paint	5 wood	1	2	3	4	5
3 A horse —	1 tail	2 harness	3 shoes	4 stable	5 rider	1	2	3	4	5
4 A train —	1 windows	2 passengers	3 wheels	4 iron doors	5 diner	1	2	3	4	5
5 An orchestra —	1 hall	2 conductor	3 drum	4 instruments	5 audience	1	2	3	4	5
6 A game —	1 players	2 cards	3 tables	4 penalties	5 goals	1	2	3	4	5

Specific Intelligence Tests. In contrast with the general intelligence tests which attempt to measure broadly the ability to learn are the tests of specific intelligence which attempt to measure ability to learn in relatively narrow fields of subject matter or areas of performance.

Aptitude Tests. Aptitude tests are frequently referred to as tests of special intelligence. They attempt to measure the aptitude of a person, and to forecast his probable future success, in certain school subjects or certain areas of performance. They are designed for use with persons who may or may not have had previous experience in the achievement areas with which they deal. Such tests attempt to measure the potentialities for success apart from those abilities resulting from specific training. Aptitude tests are not necessarily used for predictive purposes, although that is probably their most common use.

Specific intelligence refers to intelligence applied to a narrow area of performance. Thus, aptitude tests are found for such areas as English, foreign languages, music, art, mathematics, sciences, etc., and for such specific subjects as algebra, geometry, physics, and chemistry. The accompany-

⁷ Rudolf Pintner, *Pintner General Ability Tests, Verbal Series*, Intermediate. Published by World Book Co., 1938.

ing excerpts from the *Iowa Algebra Aptitude Test* illustrates the number series type of item rather common to aptitude tests in mathematics. It is apparent that many persons who could perform the necessary arithmetic operations for answering item 42, e.g., $3\frac{3}{4} \times 3 = 11\frac{1}{4}$, and $11\frac{1}{4} \times 3 = 33\frac{3}{4}$, would not do so because they failed to discover the "pattern" of the number series, in which each number is three times as large as its predecessor.

EXCERPTS FROM IOWA ALGEBRA APTITUDE TEST⁸

Part 3. NUMERICAL SERIES

Time allowance — 12 minutes

Directions: Each of the following number series is made up according to some rule. Multiplication, division, addition, and subtraction, and various combinations of these processes are employed in forming the different series. Discover the rule for each example, and write the next *two* terms on the two blank lines.

Sample: 1 2 3 4 5 6 7 8

The next two terms were 7 and 8, as we were counting by ones in this problem, thus 7 and 8 were put on the two blank lines. Work the following examples in a similar manner.

- | | | | | | | | | | |
|-----|-----------------|----------------|----------------|---------------|----------------|---|--|--|--|
| 2. | 9 | 8 | 7 | 6 | 5 | | | | |
| 3. | 1 | 1 | 5 | 5 | 9 | 9 | | | |
| 41. | 1 | 2 | 4 | 12 | 36 | | | | |
| 42. | $\frac{5}{108}$ | $\frac{5}{36}$ | $\frac{5}{12}$ | $\frac{5}{4}$ | $3\frac{3}{4}$ | | | | |

Readiness Tests. Reading readiness tests have for some years been used with primary school children in order to determine whether or not they have reached a level of maturity necessary for success in reading. Arithmetic readiness tests have been devised more recently for use in determining whether pupils have sufficient mental maturity to permit efficient learning of various arithmetic skills. Although there might be some question concerning the classification of readiness here, it seems that particularly for children entering school for the first time these tests more largely measure special mental abilities than the results of learning.

⁸ H. A. Greene and A. H. Piper, *Iowa Algebra Aptitude Test*. Published by Bureau of Educational Research and Service, University of Iowa, 1931.

Performance Tests. This term usually designates tests for which motor or manual responses rather than verbal or written responses are required of the pupil. They frequently involve pantomime rather than verbal or printed directions. They are usually at a rather low level of difficulty. They are often devised for use with illiterates, backward children, persons who are unfamiliar with English although they may read and speak a foreign language, and handicapped persons of various types.

Performance tests are of several types, which cut across the classifications of intelligence tests given above. Some are individual and others are group tests. Some measure general intelligence and others measure special aptitudes. One type is illustrated by the accompanying sample items from a revision of the *Army Beta*, a group, paper-and-pencil test of general intelligence requiring no handwriting, given to illiterates and others not able to read and speak English with ease. Another type of general intelligence test given to one person at a time requires such performances as fitting of blocks into form boards, putting together of what resembles a jig-saw puzzle, imitating actions of the examiner, etc. Still others, making use of manipulative tests similar to the above except that they require more dexterity and place a premium upon speed of response, are individual tests used in the measurement of certain types of mechanical aptitude for adolescents and even adults.

IV. PERSONALITY INSTRUMENTS

Although psychologists are in agreement that the common conception of personality is not psychologically sound, they are not in agreement concerning the real meaning of the term. They do, however, believe that personality has to do with the total behavior of the individual, both that which can and that which cannot be observed.

Measurement of personality is of recent origin. Personality instruments reflect variously the ideas of the persons constructing them. In the discussion which follows, three of the types of behavior generally classified under personality which seem to be most useful concepts to the teacher are dis-

EXCERPTS FROM KELLOGG-MORTON REVISED BETA EXAMINATION⁹

TEST 2

Put the right number under every mark




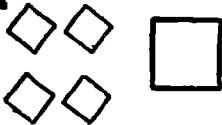
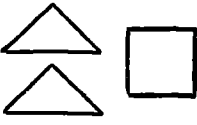
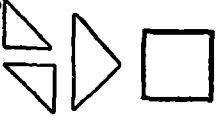
-	И	□	L	U	0	Λ	X	=
1	2	3	4	5	6	7	8	9

И	-	□	-	И	-	□	И	-	И	□	И	-	□	-
2	1	3												

L	И	□	И	L	□	-	U	И	-	L	□	-	U	L

TEST 4

Mark each square to show how the pieces at its left will fit into it

1 	2 	3 
4 	5 	6 

cussed. These, as well as some other types of behavior usually listed under personality, are discussed more completely in Chapter XI.

Attitudes Scales. The attention which has recently been called to attitudes by several nation-wide surveys of public opinion illustrates the educational importance of attitudes. Attitudes are formed, crystallized, and sometimes modified

⁹ C. E. Kellogg and N. W. Morton, *Revised Beta Examination*. Published by The Psychological Corporation, 1935.

or changed in the home, the church, on the playground, and elsewhere, as well as in the school.

Attitudes scales are of several types, but they frequently are based on a two-, three-, or five-point scale of agreement-disagreement with statements concerning controversial issues or at least issues on which opinions may readily differ. Some such scales deal with a specific issue, such as attitude toward the Chinese. Others are generalized, and may deal equally well with attitudes toward any racial group, or some other general quality.

The results from the measurement of attitudes are useful in a variety of ways both in school and in social situations, for certainly attitude changes occur as one type of instructional outcome and pupil attitudes undoubtedly influence their adjustment in the school.

Interests Inventories. The interests of different individuals vary tremendously. Not only are the individual's fields of interest sometimes obscured intentionally or unintentionally by his behavior, but in some instances his real interests may be unknown to him. Interests questionnaires use techniques somewhat similar to those for attitudes testing, and frequently request indications of the presence of interest or the degree of interest a person has in various occupations, modes of behavior, types of activity, etc.

Results from interests inventories are rather widely used in vocational guidance, and also have uses for the teacher in aiding him to adapt his instruction to pupil interests.

The accompanying illustration from one of the interest parts of the Pressey *Interest-Attitude Test* shows one method of measuring interests in a variety of things.

Adjustment Inventories. Adjustment inventories attempt to measure emotional adjustment primarily. Known by a variety of names—personality tests, personality inventories, personality schedules, adjustment inventories, and in various other ways—they ask the pupil to respond objectively to items probing his behavior, his likes and dislikes, his environment, and many other aspects of his life. A major purpose of such instruments is to locate those abnormalities and peculiarities of behavior, neurotic tendencies, etc., which should receive immediate attention if the individual possessing them is to become a well-adjusted adult.

EXCERPTS FROM PRESSEY INTEREST-ATTITUDE TEST, TEST III¹⁰

Directions. Below is a list of things that people often like or are interested in. Place a cross (X) on the dotted line in front of everything which YOU like or in which YOU are interested. Place two crosses (XX) in front of everything in which you are VERY MUCH interested. . . which you like VERY MUCH. You may mark as many or as few words as you wish. But be sure to mark everything which you like or in which you are interested.

- | | | |
|------------------------|--------------------------|----------------------------|
| 1. artist | 31. card parties | 61. dress |
| 2. drawing | 32. dancing | 62. reading |
| 3. cartoonist | 33. doctors | 63. children |
| 4. movie star | 34. fashions | 64. professors |
| 5. engineers | 35. leaders | 65. science |
| 6. comedies | 36. photography | 66. studying |
| 7. riding | 37. poker | 67. social affairs |
| horseback | 38. society | 68. coffee |
| 8. soldiers | 39. university | 69. cards |
| 9. typewriting | 40. auto driving | 70. waltzes |
| 10. carnival | | |

The accompanying excerpts from the *Student Form* of the Bell *Adjustment Inventory* illustrate items dealing with the (a) home, (b) health, (c) social, and (d) emotional adjustment of the individual.

EXCERPTS FROM BELL ADJUSTMENT INVENTORY, STUDENT FORM¹¹

DIRECTIONS

Are you interested in knowing more about your own personality? If you will answer *honestly* and *thoughtfully* all of the questions on the pages that follow, it will be possible for you to obtain a better understanding of yourself.

There are *no right or wrong* answers. Indicate your answer to each question by drawing a circle around the "Yes," the "No," or the "?". Use the question mark only when you are certain that you cannot answer "Yes" or "No." There is no time limit, but work rapidly.

If you have *not* been living with your parents, answer certain of the questions with regard to the people with whom you have been living.

- 14 Yes No Do you day dream frequently?
- 15 Yes No Do you take cold rather easily from other people?
- 16 Yes No ? Do you enjoy social gatherings just to be with people?
- 17 Yes No ? Does it frighten you when you have to see a doctor about some illness?
- 18 Yes No ? At a reception or tea do you seek to meet the important person present?
- 19 Yes No ? Are your eyes very sensitive to light?
- 20 Yes No ? Did you ever have a strong desire to run away from home?
- 21 Yes No ? Do you take responsibility for introducing people at a party?
- 22 Yes No ? Do you sometimes feel that your parents are disappointed in you?
- 23 Yes No ? Do you frequently have spells of the "blues"?

¹⁰ S. L. Pressey, *Interest-Attitude Test*. Published by The Psychological Corporation, 1933.

¹¹ Hugh M. Bell, *The Adjustment Inventory*, Student Form. Published by Stanford University Press, 1934.

TOPICS FOR DISCUSSION

1. Distinguish the three general types of tests — educational, intelligence, and personality. What are mental tests?
2. Distinguish between tests and scales. What are scaled tests?
3. Indicate the major differences between rate and power tests.
4. Briefly characterize the three forms of educational tests — oral quiz, essay examination, and objective test.
5. Indicate the major characteristics of survey and prognostic tests of general achievement.
6. Distinguish clearly between diagnostic and analytic tests.
7. In what major respects do quality and product scales differ? In what respects are they similar?
8. For what subject fields are quality scales and product scales most widely used? Why aren't tests used in these situations rather than scales?
9. Give a few illustrative essay questions and comment on their characteristics.
10. Illustrate several types of items used in informal objective tests.
11. Distinguish between tests of general intelligence and of specific intelligence. What do intelligence tests measure?
12. Briefly note some of the differences between individual and group tests of general intelligence.
13. What do aptitude and readiness tests measure? For what subject fields are they provided?
14. Briefly indicate the major characteristics and uses of performance tests.
15. Briefly characterize attitudes scales, interests inventories, and adjustment inventories. What are their major uses?

SELECTED REFERENCES

- Bingham, Walter V, *Aptitudes and Aptitude Testing*. New York · Harper and Brothers, 1937.
- Boynton, Paul L, "Intelligence and Intelligence Tests" *Encyclopedia of Educational Research*, pp. 622-34. New York · The Macmillan Co., 1941
- Freeman, Frank N, *Mental Tests Their History, Principles, and Applications* (Revised Edition). Boston · Houghton Mifflin Co., 1939.
- Fryer, Douglas, *The Measurement of Interests*. New York · Henry Holt and Co., 1931
- Hull, Clark, *Aptitude Testing*. Yonkers-on-Hudson, N. Y. World Book Co., 1928
- Hunt, Thelma, *Measurement in Psychology*. New York · Prentice-Hall, Inc., 1936.
- Lang, Albert R, *Modern Methods in Written Examinations*, Chapters IV-V Boston · Houghton Mifflin Co., 1930.

- Lee, J. Murray, *A Guide to Measurement in Secondary Schools*, Chapter II New York D Appleton-Century Co, Inc, 1936
- Lincoln, Edward A, and Workman, Linwood L, *Testing and the Use of Test Results*, Chapter II. New York The Macmillan Co, 1935.
- Nelson, M J., *Tests and Measurements in Elementary Education*, Chapter II. New York The Cordon Co, 1939
- Odell, C. W, *Traditional Examinations and New-Type Tests* New York The Century Co, 1928
- Orleans, Jacob S, *Measurement in Education*, Chapters 4-5. New York: Thomas Nelson and Sons, 1937
- Pintner, Rudolf, *Intelligence Testing* (New Edition), Chapters VI-VII. New York Henry Holt and Co, 1931
- Remmers, H H, and Silance, E. B, "Generalized Attitude Scales." *Journal of Social Psychology*, 5 298-312, August 1934
- Rinsland, Henry D, *Constructing Tests and Grading in Elementary and High School Subjects* New York Prentice-Hall, Inc, 1938.
- Ruch, G. M, *The Objective or New-Type Examination*. Chicago: Scott, Foresman and Co, 1929
- Russell, Charles, *Classroom Tests* Boston Ginn and Co, 1926
- Russell, Charles, *Standard Tests* Boston Ginn and Co, 1930.
- Stagner, Ross, "Attitudes." *Encyclopedia of Educational Research*, pp. 69-75. New York The Macmillan Co, 1941
- Symonds, Percival M., *Diagnosing Personality and Conduct*, Chapters V-IX New York . D. Appleton-Century Co., Inc., 1931.
- Terman, Lewis M, and Merrill, Maud A, *Measuring Intelligence*. Boston Houghton Mifflin Co, 1937
- Thorpe, Louis P, *Psychological Foundations of Personality*, Chapters I, XI. New York McGraw-Hill Book Co, Inc, 1938
- Thurstone, L L, and Chave, E. J., *The Measurement of Attitude* Chicago University of Chicago Press, 1929
- Webb, L W, and Shotwell, Anna Markt, *Testing in the Elementary School*, Chapter III. New York Farrar and Rinehart, Inc., 1939.

CHAPTER III

DEVELOPMENT OF EDUCATIONAL AND MENTAL TESTING

The discussion of this chapter traces educational and mental testing from the time of the earliest historical records up to the present :

- a.* Measurement to 1800.
- b.* Educational testing during the nineteenth century.
- c.* Intelligence testing during the nineteenth century.
- d.* Development of modern intelligence testing.
- e.* Development of modern educational testing.
- f.* Development of modern personality testing.
- g.* Present status of educational and mental measurements.

Measurement of human behavior with primary reference to the educational attainments and capacities of school children can well be divided roughly into three periods. During the first period, from the beginning of historical records down to about the nineteenth century A.D., educational measurements were naturally quite crude. Although the fact that individuals differ widely in their capacities and abilities has been recognized for several thousand years and educational measurement made formal entrance to the schools as early as medieval times, relatively little progress in educational testing was made until the opening of the present century. During the second period, embracing approximately the nineteenth century, educational measurement began to assimilate from various sources the ideas and the scientific and statistical techniques which were later to result in the modern objective testing movement. The brief third period, dating from about 1900 to the present, is characterized by tremendous advances in statistical techniques, in the measurement of achievement, intelligence, and personality, and, during the last decade, by increased clarity of thinking concerning the nature of ability, the tools used in its measurement, and the proper use and further improvement of those tools.

I. MEASUREMENT TO 1800

Early Oral Examinations. The first evidences of the oral examination are found in ancient literature. The story is told in the Old Testament (Judges, 12 5-7) of the test the Gileadites gave to the enemy Ephraimites who wished to cross the Jordan. When asked to pronounce the word "Shibboleth," the Ephraimites could answer only with "Sibboleth," whereas people of the friendly tribes could respond with the correct pronunciation. Forty-two thousand Ephraimites were killed because they failed to pass this objective test.¹ Socrates, in a method he made famous, subjected his pupils to exhaustive and searching questioning. Oral quizzing, Socratic or otherwise, has undoubtedly been a part of classroom procedure from the beginnings of teaching activity—in fact, there have been and still are times when, for certain teachers, it constitutes practically the whole of the teaching act.

Early Written Examinations. Written tests are probably of more recent origin than oral quizzes, but even they date back many centuries. As early as 2200 B.C., China had an elaborate national system of examinations for the purpose of selecting her public officials, and these examinations have been known down through the ages for their unusual severity. Confined in isolated cells for hours at a time, candidates were compelled to write lengthy papers or treatises on assigned topics.²

Recognition of Individual Differences. Individual differences among people have long been recognized. Plato, nearly four centuries B.C., divided his ideal society into the three classes: (1) workers, (2) protectors, and (3) rulers. He believed that persons suited to each class should receive education for the fullest development of their personalities.³ Quintillian, shortly after the start of the Christian era, wrote that masters should observe differences in ability and inclina-

¹ Norma V Scheidemann, "The Earliest Recorded Objective Test" *School and Society*, 20 702, June 1, 1929

² W A P Martin, *The Chinese Their Education, Philosophy, and Letters*, pp. 45-49 Harper and Brothers, New York, 1881

³ Edgar W. Knight, *Twenty Centuries of Education*, p. 62. Ginn and Co., Boston, 1940.

tions of persons they instructed, for the "forms of mind are not less varied than those of bodies."⁴

First Educational Tests. The first tests used for the measurement of the results or outcomes of education were probably not unlike certain of the performance tests of today, at least to the extent that they measured physical performance and that they were not paper and pencil tests.

Among various primitive tribes, in which the young men were taught to hunt, fish, and fight, the initiation ceremonies prerequisite to their admission to the ranks of adult males tested knowledge of tribal customs, endurance, bravery, and other knowledges and abilities thought necessary for tribal protection.⁵

The ancient Spartans, whose educational curricula for their youth stressed physical development and stoicism, conducted examinations as early as 500 B.C. in which the young men underwent painful ordeals.⁶ In ancient Athens, the stress upon athletics and aesthetic development led to evaluation by means of games and contests and of reading, writing, and singing ability.⁷

First Tests in the School. In medieval times, the oral examination was used in universities. The University of Bologna by 1219 A.D. and the University of Paris before the close of the thirteenth century required degree candidates to defend their theses orally. However, the written educational examination probably made its first appearance for educational use at Cambridge, England, in 1702.⁸

II. EDUCATIONAL TESTS FROM 1800 TO 1900

Early Educational Tests in America. According to available records,⁹ the first examinations of note in this country

⁴ William Boyd, *The History of Western Education*, p 76 A and C Black, Ltd, London, 1921

⁵ Charles Russell, *Standard Tests*, pp 14-15 Ginn and Co, Boston, 1930.

⁶ *Ibid* p 16

⁷ Knight, *op cit* pp 52-53.

⁸ Albert R. Lang, *Modern Methods in Written Examinations*, pp 2-3 Houghton Mifflin Co, Boston, 1930

⁹ Otis W Caldwell and Stuart A Courtis, *Then and Now in Education, 1845-1923*, Chapters I, III. World Book Co, Yonkers-on-Hudson, N. Y., 1923.

were those of Boston in 1845. One of the specified duties of the Boston school committee was to make an inventory of the schools each year. This annual inspection included an oral examination of all the pupils. As the number of pupils increased, this practice naturally became impossible. It was then decided to quiz only the highest class in each school. This soon became impossible also because of the rapid growth in the number of pupils, and as a result the examining became quite perfunctory. Finally, the sub-committee, which was appointed to survey the grammar departments of the Boston schools in 1845, decided to use written examinations. These examinations covered the subjects of arithmetic, astronomy, geography, grammar, history, and natural philosophy. The details of preparation, administration, and interpretation of results were intended to make the examinations as fair as possible. Questions were carefully graded in terms of difficulty, and scoring rules were prepared. On the basis of these examinations, the schools were ranked in order of merit. The sub-committee apparently recognized the problem of "immeasurables," as they emphasized the fact that the tests measured intellectual activity and acquirements only, and that inclusion of traits of conscience, respect for order, religious sense, and the like might modify the ranking of the schools.¹⁰

This Boston examination project is truly a high light in the history of education in the United States. It made a real impression upon Horace Mann, who at that time was Secretary of the Massachusetts Board of Education.¹¹ As editor of the *Common School Journal*, he published extracts from the report and made many noteworthy comments¹² on the subject of examinations. He concluded that the new written examination was so superior to the old oral quiz that no school committee would ever again venture to relapse into the former inadequate and uncertain practice. The reasons ad-

¹⁰ Ibid p 181 From report of Grammar School Committee to the School Committee of the City of Boston, May 6, 1845, Theophilus Parsons, S G Howe, and Rollin H Neale, members

¹¹ Horace Mann doubtless exerted considerable influence on the sub-committee, so the examinations were probably reflections of his ideas

¹² Horace Mann, "Boston Grammar and Writing Schools" *Common School Journal*, Vol VII, No 19, October 1, 1845 Also reported in Caldwell and Courtis, op cit. pp 237-72.

vanced by Horace Mann in support of the written examination were as follows:¹³

1. It is impartial.
2. It is just to the pupils.
3. It is more thorough than older forms of examination.
4. It prevents the "officious interference" of the teacher
5. It "determines, beyond appeal or gainsaying, whether the pupils have been faithfully and competently taught"
6. It takes away "all possibility of favoritism."
7. It makes the information obtained available to all
8. It enables all to appraise the ease or difficulty of the questions.

Although disagreement is possible with the accuracy of this characterization of the written examination, it must be admitted that his ideals were apparently those represented by modern tests, but his instruments were inadequate. It is significant to note also that in successive issues of the *Common School Journal* Mann suggested most of the elements in examinations which are found in the modern measurement movement.

Early Objective Tests. To Rev. George Fisher, an English schoolmaster, goes the credit for devising and using what were probably the first objective measures of achievement. His "scale books" were in use in the Greenwich Hospital School as early as 1864. Provided for handwriting, spelling, mathematics, navigation, Scripture knowledge, grammar and composition, French, general history, drawing, and practical science, they scaled performance by units of one-fourth from 1, representing the highest, to 5, representing the lowest, degrees of efficiency. In such subjects as handwriting and drawing, where qualitative rather than quantitative evaluation was the custom, specimens of the pupil's work were compared with "standard specimens" to determine the pupil's numerical rating. The numerical values for spelling and other subjects to which quantitative measures of achievement were commonly applied depended upon percentages of errors in performance.¹⁴

¹³ Caldwell and Curtis, op cit p 37

¹⁴ E B Chadwick, "Statistics of Educational Results" *The Museum, A Quarterly Magazine of Education, Literature, and Science*, 3, 480-84, January 1864. Also in "Educational Measurements of Fifty Years Ago" *Journal of Educational Research*, 4, 551-52; November 1913

Fisher's "scale books" were doubtless somewhat crude in organization, but they included the germ of many of the ideas which are incorporated in our present-day educational scales. His work, like that of most pioneers, produced no lasting results because he lived too far in advance of the thought and educational practice of his day. His work appears to be isolated from the development of educational tests in this country.

First Objective Tests in America. In America, the real inventor of the comparative test was Dr. J. M. Rice, who in 1894¹⁵ hit upon the idea which he so effectively developed that it became the foundation of objective measurement in education. Rice, having administered a list of spelling words to pupils in many school systems and analyzed the results, confounded the educators at the 1897 session of the Department of Superintendence of the National Education Association with the declaration that pupils who had studied spelling thirty minutes a day for eight years were not better spellers than children who had studied the subject fifteen minutes a day for eight years. Rice was attacked and reviled for this "heresy," and some educators even attacked the use of a measure of how well pupils could spell for evaluating the efficiency of spelling instruction. They contended that spelling was taught to develop the pupils' minds and not to teach them to spell. Although Rice continued his work effectively and educators gradually rallied to his support, it was more than ten years later that his pioneering resulted in significant attention to the objective method in educational testing.¹⁶

III. INTELLIGENCE TESTING FROM 1800 TO 1900

Scientific Recognition of Individual Differences. It was not, apparently, until 1796, that individual differences in mental abilities were first brought under not the microscope, but, literally, the telescope. It was in that year at the

¹⁵ Leonard P. Ayres, "History and Present Status of Educational Measurements" *The Measurement of Educational Products* Seventeenth Yearbook of the National Society for the Study of Education, Part II, Chapter I, p. 11. Public School Publishing Co., Bloomington, Ill., 1918.

¹⁶ *Ibid* p. 12.

Greenwich Astronomical Observatory in England that one of the observers who recorded the instant of time at which stars crossed the lines on telescope lenses was discharged because his observations consistently differed slightly from those of his colleagues. In 1816, however, it was discovered by an astronomer who read an account of this incident that an error of observation, called the "personal equation," characterized the work of all observers and that the amount of error varied from person to person and also in the same person from time to time.¹⁷ As a result, by 1822 astronomers were recognizing and allowing for this difference among observers in their reaction time.

Scientific Study of Individual Differences. Galton, with the publication of his *Hereditary Genius* in 1869, brought the scientific study of individual differences into focus, developed it further by instituting measurement of various human physical traits and motor abilities, and even investigated mental ability by methods which many years later became highly fruitful.¹⁸

Foundations of Statistical Method. Galton's most important contribution to educational measurements was not in the field of individual differences, however, but in the derivation of statistical methods. Here, in devising a system of "standard scores" and in developing graphically the idea for an objective measure of relationship, the correlation coefficient, he furnished tools essential not alone to the development of educational and mental testing but also to scientific method in education. Pearson later formulated the method now most commonly used for calculating the correlation coefficient.¹⁹

Early Attempts to Measure Intelligence. Dr. E. S. Chaille, an American physician, is credited as early as 1887 with the development of standards and simple tests for judging the mental levels of children to the age of three and

¹⁷ Anne Anastasi, *Differential Psychology*, pp. 9-10. The Macmillan Co., New York, 1937.

¹⁸ Joseph Peterson, *Early Conceptions and Tests of Intelligence*, pp. 73-75. World Book Co., Yonkers-on-Hudson, N. Y., 1925.

¹⁹ Henry L. Garrett, *Great Experiments in Psychology*, pp. 171-74. The Century Co., New York, 1930.

with having implied, although not definitely used, the concept of mental age.²⁰

Cattell apparently first used the term "mental test"²¹ in 1890, almost at the beginning of the period during which scientific method was first applied to the measurement of mental ability. Cattell, Wissler, and Jastrow were prominent among the American experimenters devoting attention to intelligence during the last decade of the nineteenth century. Attempts to measure intelligence by means of physical characteristics, and sensory acuity and motor skills tests gave, for the most part, negative results.²²

During the same period, Binet and his colleagues were experimenting in France with tests of a somewhat similar but less specific type. In 1895, Binet and Henri described ten types of tests which, differing from American tests mainly in the much greater complexity of behavior they would measure, they thought were likely to discriminate between levels of mental ability.²³

IV. INTELLIGENCE TESTS FROM 1900 TO THE PRESENT

First Individual Intelligence Tests. Binet and Simon brought out the first intelligence scale in 1905, devising it primarily for the purpose of selecting mentally retarded pupils who required special instruction. This pioneer individual intelligence scale consisted of thirty parts dealing with widely varying abilities, ranging in order from very easy to very difficult. It utilized the basic idea of interpreting the relative intelligence of different children at any given chronological age by the number of tests they could pass. These characteristics were all reembodyed in the 1908 and 1911 Revisions of the *Binet-Simon Scale*, published by the original authors, and also are basic to most individual intelligence scales even today. The 1908 Revision introduced the funda-

²⁰ Florence L. Goodenough, "An Early Intelligence Test" *Child Development*, 5 13-18, March 1934

²¹ J. McKeen Cattell, "Mental Tests and Measurements" *Mind*, 15 375-81, July 1890

²² Frank N. Freeman, *Mental Tests Their History, Principles, and Applications* (Revised Edition), p. 58. Houghton Mifflin Co., Boston, 1939

²³ Anastasi, *op cit* p. 19

mentally important concept of mental age (MA) and provided means for obtaining it.²⁴

Individual Intelligence Tests in America. Goddard, Kuhlmann, and Terman all adapted the Binet-Simon tests to use with American children during the period from 1911 to 1916. Terman and his collaborators made the *Stanford Revision of the Binet Scale* available in 1916, and in 1937 followed it with a second and more complete revision. These revisions, which have been the most widely used of any in America, make use of the intelligence quotient (IQ), based on the relationship between a child's mental age and his chronological age.²⁵

First Group Intelligence Tests. Although various psychologists had been working on a group intelligence test, and Otis was near the point of issuing such a test around 1917, the *Army Alpha* test, used for measuring and placing American Army recruits and draftees during the First World War, was the first group intelligence test to be published. The *Army Alpha* test, which was widely used for testing men who could read and understand English, was accompanied by *Army Beta*, a non-language test for use with illiterates and men who, although perhaps literate in a foreign language, could not read English. Both tests were the work of a group of psychologists, including Yerkes, Yoakam, Whipple, and Otis, whose offer of service was accepted by the U. S. Army.²⁶

Later Development of Group Intelligence Tests. Group intelligence tests began making their appearances almost immediately following the end of the First World War, and the period from 1918 to the middle twenties was marked both by the publication of many such tests and by an upsurge of general interest in intelligence testing. Although the testing techniques have been refined considerably in the past decade, and a few tests of recent publication recognize by part scores the two factors of linguistic and quantitative thinking as components of mental ability, the past fifteen

²⁴ Freeman, op cit. pp. 86-88.

²⁵ Ibid p 101

²⁶ Ibid pp 113-14.

years have brought no outstanding changes in the methods of measuring general intelligence.

Aptitude or Specific Intelligence Tests. The measurement of aptitudes, or those potentialities for success in an area of performance which exist prior to direct acquaintance with that area, has been tied up with intelligence testing both fore and aft. Early attempts to measure general intelligence were by means of tests of many specific traits and aptitudes, but that approach was dropped when Binet showed that tests of more complex forms of behavior were superior. It was soon apparent, however, that general intelligence tests were not highly predictive of certain types of performance, especially in the trades and industries.

Münsterberg's aptitude tests for telephone girls and street-car motormen in 1913 were followed by tests of mechanical aptitude, musical aptitude, art aptitude, clerical aptitude, and aptitude for various subjects of the high school and college curricula prior to 1930.²⁷ Spearman's splitting of total mental ability into a general factor and many specific factors²⁸ had its influence on this movement, and accounted for the fact that aptitude tests are frequently called specific intelligence tests.

V. EDUCATIONAL TESTS FROM 1900 TO THE PRESENT

First Book on Educational Measurement. Thorndike brought out the first book dealing primarily with mental and educational measurements in 1904,²⁹ and both through this book and his later influence on his students became more than any other person responsible for the early development and popularization of standardized educational tests.

First Standardized Achievement Tests. Stone, a student of Thorndike's, published his arithmetic reasoning test, the

²⁷ Goodwin Watson, "The Specific Techniques of Investigation Testing Intelligence, Aptitudes, and Personality" *The Scientific Movement in Education* Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XXX, pp 365-66 Public School Publishing Co., Bloomington, Ill., 1938

²⁸ Charles Spearman, "'General Intelligence' Objectively Determined and Measured" *American Journal of Psychology*, 15 201-93, April 1904

²⁹ Edward L. Thorndike, *An Introduction to the Theory of Mental and Social Measurements*. Teachers College, Columbia University, New York, 1904

first standardized instrument to make its appearance, in 1908.³⁰ Thorndike in 1909 published his *Scale for Handwriting of Children*—the first standardized achievement scale.³¹ During the period 1909 to 1915, Courtis published a series of arithmetic tests, while five scales for measuring abilities in English composition, spelling, drawing, and handwriting were published by Hillegas, Buckingham, Thorndike, and Ayres.³² It is interesting to note that only two of these pioneer instruments were tests, while the remaining five were scales.

Educators at first opposed the standardized test and derided the testers. Although the Department of Superintendence of the National Education Association at its 1914 meeting supported the report of a Committee on Tests and Standards, it had voted against measurement after a heated discussion two years earlier.³³ However, the spread of standardized testing continued, under the stimulation of at least three early developments:

(1) The numerous important studies of the accuracy of school marks revealed the fact that school marks are highly subjective, and hence inaccurate. This accumulation of evidence was conclusive proof not only that such subjective measures as teachers' marks might be expected to vary widely from teacher to teacher but also that the same teacher's marks would vary from time to time and from subject to subject. It was shown that very often the difference between success and failure in a given subject was largely the pupil's fortune in being placed in one teacher's section of the class rather than in that of another. This type of demonstration revealed the need for instruments which would yield more accurate measures of achievement.

(2) The surveys or investigations of certain of the larger school systems, resulting from a prevailing feeling of dissatisfaction with existing conditions and coupled with a desire on

³⁰ Cliff W Stone, *Arithmetical Abilities and Some Factors Determining Them*. Contributions to Education, No 19. Teachers College, Columbia University, New York, 1908

³¹ Edward L Thorndike, "Handwriting" *Teachers College Record*, 11 83-175, March 1910

³² C W Odell, *Educational Measurements in High School*, pp. 34-35. The Century Co., New York, 1930

³³ Ayres, op. cit p 14

the part of a few educators to reveal true conditions, both stimulated the construction and use of tests and were influenced by the derivation of more objective devices for measuring the abilities of pupils.

(3) The development of educational measurements was greatly facilitated by the establishment of research organizations. Research bureaus were organized in many of the larger school systems primarily for the purpose of directing and stimulating the application of educational tests. Similar research bureaus have been established in a large number of state educational institutions, and recognition has been given to the work by the department of public instruction in a number of states. Such types of research organizations have been very influential in popularizing the use of educational tests.

Although the pioneer and most of the early standardized tests were for use in the elementary school, it was not many years until the high school and even the college were well provided with such instruments. The popularity and use of tests spread rapidly once they had been accepted as a worthwhile educational tool. However, a reaction against standardized tests, caused largely through their improper use by inexperienced and untrained persons, occurred some fifteen years after the first tests were published.

Later Development of Standardized Achievement Tests. The history of achievement measurement since the late twenties has been characterized mainly by increasing recognition of the fact that test results offer only one, although the major one, of the types of acceptable evidence on pupil achievement. This tendency toward evaluation, which is broader in scope than testing, has been accompanied by a strong trend toward more scientific use of measurement tools. The more scientific approach to measurement and the effective use of results has quelled the fears of those who looked with skepticism some years ago on the testing movement.

Origin and Development of Informal Objective Examination. The idea of the informal objective examination, referred to during its early days rather loosely as the "New-Type Test" and the "Objective Test," apparently was first publicly expressed by McCall, whose article in 1920 first sug-

gested that teachers did not need to depend solely upon standardized tests but that they could construct their own objective tests for classroom use.³⁴ The pioneer book dealing almost entirely with this testing adaptation was published in 1924.³⁵ The informal objective test has since come into such wide use that a survey in 1936 of testing practices among 1600 high school teachers widely distributed throughout the country showed that 74 percent used the informal objective and an additional 10 percent used a combination of the informal objective and essay examinations.³⁶

Although the contributions of Tyler have been significant in both the standardized testing and the informal objective testing movements, it is probably in the latter field that his influence has been the greater. He outlined steps of procedure for test construction and validation which clearly pointed out the essential dependence of a program of achievement testing upon the objectives of instruction and the recognition of forms of pupil behavior indicating attainment of the desired instructional outcomes.³⁷ Perhaps he more than any other single test specialist was responsible for the extension of achievement testing to the more intangible outcomes of instruction, for his contributions nearly ten years ago doubtless did much to bring into being the broad modern conception of evaluation to replace the earlier and narrower concept of testing.³⁸

The Progressive Education Association is now well along with an eight-year evaluation study under the general direction of Tyler as part of an investigation of the relationships between colleges and secondary schools.³⁹ The Evaluation Staff of the Commission on the Relation of School and College of the Progressive Education Association has been work-

³⁴ William A. McCall, "A New Kind of School Examination" *Journal of Educational Research*, 1 33-46, January 1920

³⁵ G. M. Ruch, *The Improvement of the Written Examination*. Scott, Foresman and Co., Chicago, 1924

³⁶ J. Murray Lee and David Segel, *Testing Practices of High-School Teachers*. U. S. Office of Education Bulletin, 1936, No. 9, p. 6. Government Printing Office, Washington, D. C., 1936

³⁷ Ralph W. Tyler, "A Generalized Technique for Constructing Achievement Tests" *Educational Research Bulletin*, 8 199-208, April 15, 1931

³⁸ Ralph W. Tyler, *Constructing Achievement Tests*. Ohio State University, Columbus, Ohio, 1934

³⁹ Ralph W. Tyler, "Appraising Progressive Schools" *Educational Methods*, 15 412-15, May 1936

ing since 1936 on a wide cooperative basis in some thirty member schools. Theirs is undoubtedly the broadest coordinated attack upon evaluation and appraisal of all types of instructional outcomes by the use primarily of informal objective examination procedures and related techniques that has ever been undertaken.

VI. PERSONALITY TESTS FROM THEIR ORIGINS TO THE PRESENT

Antecedents of Modern Personality Tests. Personality testing had its antecedents in the work of Kraepelin and Sommer on free association tests during the last decade of the nineteenth century. However, the questionnaire and rating scale methods used by Galton and others for different purposes at much earlier dates became the dominant early methods of personality measurement in America.⁴⁰

Modern Personality Inventories and Questionnaires. Woodworth devised a *Personal Data Sheet*, in reality an inventory of neurotic tendencies and emotional maladjustment, for use with American soldiers during the First World War. This was probably the outstanding early contribution in this field. During the past two decades, tools for the measurement of conduct, attitudes, vocational interests, and many other phases of personality were developed. The use of anecdotal records, observation of behavior, and case studies has recently supplemented the questionnaire and rating scale in the measurement of the rather vaguely-delimited area of behavior called personality.⁴¹

VII. THE PRESENT STATUS OF EDUCATIONAL AND MENTAL MEASUREMENT

Although educational and mental measurement are still unquestionably in the developmental stages, their merits and appropriate uses are increasingly coming to be recognized. On the other hand, many of their shortcomings are thoroughly realized. The modern emphasis on the guid-

⁴⁰ Anastasi, op. cit. pp. 23-24.

⁴¹ Watson, op. cit. pp. 368-69.

ance function of the teacher and the increased familiarity of teachers with evaluation techniques have resulted more and more in a transfer of measurement functions from the specialist to the teacher and in cooperative attacks of test specialists and subject matter specialists on common problems in this field.

TOPICS FOR DISCUSSION

1. What were some of the ancient forerunners of educational tests?
2. Show how educational testing had its origins centuries before standardized and informal objective tests were developed.
3. List and evaluate rather fully the most important ideas concerning examinations expressed by Horace Mann.
4. Discuss the "scale books" developed by Rev. George Fisher and compare them with modern educational scales.
5. Why did several of the early contributions to measurement, such as those of Mann and Fisher, fail to exert rather immediate influence on measurement practices?
6. What was the significance for objective measurement and for educational research of the contributions made by Dr J M Rice?
7. What three important educational developments of the first two decades of the present century indirectly stimulated the growth of interest in educational measurements?
8. Who were the pioneers in the development of standardized educational tests? What was their influence on the measurement movement?
9. Indicate the part played by the informal objective examination in the development of educational testing.
10. Discuss the early recognition and first objective measurement of individual differences.
11. By what method did workers in the field of mental ability first attempt to measure intelligence? How successful were their attempts?
12. Discuss the contributions of Binet and Simon to the intelligence testing movement.
13. Briefly discuss the development of group intelligence testing from the First World War to the present.
14. What types of abilities did the early aptitude tests measure?
15. Discuss the early attempts to measure personality and the more recent personality inventories and questionnaires.
16. Comment upon the status of educational and mental measurements today.

SELECTED REFERENCES

- Anastasi, Anne, *Differential Psychology*, Chapter I. New York: The Macmillan Co., 1937.
- Ayres, Leonard P., "History and Present Status of Educational Measurements" *The Measurement of Educational Products*. Seventeenth Yearbook of the National Society for the Study of Education, Part II, Chapter I, pp. 9-15. Bloomington, Ill.: Public School Publishing Co., 1918.
- Caldwell, Otis W., and Courtis, Stuart A., *Then and Now in Education: 1845-1923*. Yonkers-on-Hudson, N. Y.: World Book Co., 1923.
- Freeman, Frank N., *Mental Tests: Their History, Principles, and Applications* (Revised Edition). Boston: Houghton Mifflin Co., 1939.
- Garrett, Henry E., *Great Experiments in Psychology*, Chapters I-III, VIII-IX. New York: The Century Co., 1930.
- Hunt, Thelma, *Measurement in Psychology*, Chapter III. New York: Prentice-Hall, Inc., 1936.
- Lang, Albert R., *Modern Methods in Written Examinations*, Chapter I. Boston: Houghton Mifflin Co., 1930.
- McCall, William A., "A New Kind of School Examination." *Journal of Educational Research*, 1: 33-46, January, 1920.
- Odell, C. W., *Educational Measurements in High School*, Chapter II. New York: The Century Co., 1930.
- Peterson, Joseph, *Early Conceptions and Tests of Intelligence*. Yonkers-on-Hudson, N. Y.: World Book Co., 1925.
- Pintner, Rudolf, *Intelligence Testing* (New Edition), Chapters I-III. New York: Henry Holt and Co., 1931.
- Ross, C. C., *Measurement in Today's Schools*, Chapter II. New York: Prentice-Hall, Inc., 1941.
- Ruch, G. M., *The Improvement of the Written Examination*. Chicago: Scott, Foresman and Co., 1924.
- Russell, Charles, *Standard Tests*, Chapter II. Boston: Ginn and Co., 1930.
- Tyler, Ralph W., *Constructing Achievement Tests*. Columbus, Ohio: Ohio State University, 1934.
- Tyler, Ralph W., "The Specific Techniques of Investigation: Examining and Testing Acquired Knowledge, Skill, and Ability." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XXIX, pp. 341-55. Bloomington, Ill.: Public School Publishing Co., 1938.
- Watson, Goodwin, "The Specific Techniques of Investigation: Testing Intelligence, Aptitudes, and Personality" *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XXX, pp. 357-73. Bloomington, Ill.: Public School Publishing Co., 1938.

CHAPTER IV

CRITERIA OF A GOOD EXAMINATION

The following aspects of the criteria or distinguishing characteristics of a good examination are discussed in this chapter.

- a. Validity as an essential characteristic of a good examination.
- b. Curricular and statistical validity
- c. Reliability as an aspect of test validity.
- d. Methods of determining and estimating test reliability.
- e. Dependence of test reliability upon objectivity and adequacy of sampling
- f. Administrability and scorability as test criteria.
- g. Comparability important to the use of test results.
- h. Economy as a necessary consideration.
- i. Utility as a final, overall criterion of a good examination.

The selection of a standardized test or the construction of an informal objective examination or an essay examination for any type of testing situation require a careful consideration of the characteristics of a good examination. Although the criteria of a good examination can be listed and classified in many different ways, test specialists are in general agreement concerning the aspects of tests which should receive attention in selecting or constructing them. The nine criteria of a good examination discussed below undoubtedly represent the most important considerations to be taken into account

It is recommended that the student refer frequently to the discussion in Chapter XXIII on the statistical methods of determining test validity and reliability in connection with the study of those two exceedingly important criteria. An adequate understanding of these criteria depends upon both their theoretical and their statistical aspects.

I. VALIDITY

Validity is the most important characteristic of a good examination, for unless a test is valid it serves no useful func-

tion. *The validity of an examination depends upon the efficiency with which it measures what it attempts to measure.* A test must, therefore, accomplish the purpose the user has in mind for it in order that it satisfy this fundamental criterion for all testing. In fact, the uncritical acceptance of an invalid test by a teacher for performing a desired function might easily result in serious injustice to the pupils. Accordingly, teachers cannot be too careful in assuring themselves of the validity of the tests they use. For example, a teacher who used a test which measured only knowledge of facts in a course in American history would not be correct in drawing conclusions on the basis of the results concerning the abilities of her pupils to apply historical facts to the reasoned interpretation of events.

It follows also that a test must be used with pupils who possess the proper intellectual maturity and background of experience for taking the test if it is to possess validity. For example, a standard arithmetic survey test intended for use with pupils in Grades 6 to 9 might be invalid for use with most of the pupils in Grade 5 and probably with all pupils in the lower grades.

Lindquist¹ illustrates validity by pointing out that a test of high validity for ranking high school pupils on general achievement in United States history would have constantly decreasing validities for testing college students over a course in the same subject, for testing high school pupils over a course in economic history of the United States, for predicting future success in a secondary school English history course, for diagnosing weaknesses in abilities in United States history, for measuring general intelligence, and, finally, as a basis for assigning course marks in manual training.

Validity is therefore a specific rather than a general criterion of a good examination. It is specific in the sense that a test may be highly valid for use in one situation and highly invalid for use in another manner. It is specific also in the sense that a test may be valid for use with one group

¹ Herbert E Hawkes, E F Lindquist, and C R Mann (Editors), *The Construction and Use of Achievement Examinations*, pp 21-22 Houghton Mifflin Co, Boston, 1936.

of pupils but not for use with a different pupil group. *Tests cannot correctly be described as valid in general terms, but only in connection with their intended use and at the intended ability level of pupils.*

There is a difference in the concept of validity which should be applied in the consideration of standardized and informal objective examinations. It is readily apparent that the teacher better than anyone else knows the content and emphases of the course he has taught, so in that sense he is the person best qualified to construct a valid test for his course. However, it is frequently true that the makers of standardized tests are better able than many, if not most, classroom teachers to determine what commonly are, and perhaps what should be, the content and emphases in courses for which they construct and standardize tests. Therefore, it seems reasonable to conclude that insofar as test content is concerned the teacher is the person best qualified to test the attainment of the desired outcomes in his particular class, but that the standardized test affords a superior means for determining how well his pupils have attained the outcomes which are most widely recognized as being desirable in the particular course he is teaching. This difference in the application of the concept of validity is the result of the fact that no two teachers teach *exactly* the same course and that no one teacher teaches *exactly* the same course twice during his lifetime. Although this is particularly true for courses in the contemporary social studies and literature, and in the sciences, in which new content must be introduced constantly to keep abreast of developments, it is true even of such subjects as mathematics, in which the methods used and classroom problems which arise may well differ from semester to semester even though the basic content may be largely unchanged.

Three types of test validity are discussed below—curricular validity, statistical validity, and psychological and logical validity. Of these three, the first is by far the most important, for in the final analysis any method of test validation must be based on relatively subjective judgment concerning the degree to which the test covers the proper ground. Statistical validity, in turn, is a more widely used

and probably more important concept than psychological and logical validity.

Curricular Validity. The first of the three types of methods used in determining the validity of a test is curricular validation. A teacher who carefully and thoughtfully selects a standardized test or constructs an informal objective examination for his class is attempting to insure curricular validity by making certain that the test deals with the types of educational outcomes he wishes to measure and is at the proper level of difficulty for his pupils. There are various sources of evidence which can guide the teacher in considering test validity from the curricular standpoint. Among these are textbooks, courses of study, reports of national or regional committees, and the writings of subject-matter specialists. The idea in each case is that analysis of these source materials will furnish evidence concerning the thinking of qualified educators on questions dealing with course content, emphases, and methods, and that they afford the only objective basis for determining the important outcomes to test.

Textbook and Course of Study Analyses. The major fallacy in the analysis of textbook and course of study content as a validation method is that it tends to perpetuate faulty and inadequate curricular content if such is present. It does not look beyond present practices. On the other hand, the overlapping of instructional material which is common to a large number of textbooks and courses of study is almost certainly important content. Thus, if one were to attempt to validate a test in geography or history, one way to determine its content would be to include in it only those outcomes and skills which receive definite instructional emphasis in, say, five out of six courses of study or textbooks which have the confidence of a large body of school administrators.

Reports of National or Regional Committees and Writings of Subject Specialists. Although these methods may not tend to perpetuate undesirable present practices, such reports are so seldom accompanied by instructional materials at the time of their appearance that the teacher wishing to use their recommendations in considering the validity of his

tests finds no suitable standardized tests available and would have to break virgin soil in the construction of any informal objective tests which would meet his requirements.

An illustration of the application of the method of validating test material in terms of judgment of qualified authorities is found in the development of the *Iowa Silent Reading Test, New Edition, Elementary Examination*. The following tabulation shows the unit skills contributing to the pupil's ability to read and to work with books which are measured by these reading tests :²

Test 1. Rate and Comprehension

Science material

Social studies material

Test 2. Directed Reading

Science material

Social studies material

Test 3. Word Meaning

General vocabulary

Subject-matter vocabulary

Test 4. Paragraph Comprehension

Selection of central idea of paragraph

Identification of details essential to the meaning of the paragraph

Test 5. Sentence Meaning

Test 6. Locating Information

Alphabetizing : Use of guide words

Use of index

If the above list of measurable skills is compared with the major objectives of reading instruction given in Chapter XV, pages 328 to 331, the relationship will be apparent. The case for the validity of this reading examination rests on the exactness with which these objectives of reading have been paralleled in the test parts comprising it. Then the validity of the test is determined by the extent to which it measures the desirable skills in silent reading as recognized by specialists in this field. In achievement tests based more generally upon information than upon skills, the validity of the test depends more largely upon the opportunity the pupil has had to master the content measured in the test. In such

² H A Greene and V H Kelley, *Manual of Directions Iowa Silent Reading Tests, New Edition, Elementary Examination*, p. 3 World Book Co., Yonkers-on-Hudson, N Y, 1939

a situation the teacher himself is probably the best judge of the validity of the test, since he knows best what material he has taught the class.

Statistical Validity. A second method of validating tests is by means of statistical techniques. Methods frequently used involve the determination of the correlation between test scores and such criteria as teachers' marks, ratings of expert judges, scores on other tests designed for the same type of use, and measures of success on certain types of future outcomes. Basic to this method is the belief that the test is valid if high correlations are obtained between scores on it and the criterion measures, and implied is the belief that the criterion measures may be accepted as measurement standards. Correlation coefficients obtained from the types of situations named above are called *validity coefficients* or *coefficients of validity*.

Correlation with School Marks. The method of validation by correlation with school marks assumes that in the long run a test has validity if the pupils' scores on it are closely related to their achievement in the subject. That is, a test in language must have considerable validity if pupils whose school marks in the subject are consistently high make the superior scores on the test, and if pupils whose school marks in the course are low usually make the inferior scores on the test. In spite of the apparent unreliability of teachers' marks for refined measurements, it is probably true that a teacher who has had a semester or a year in which to observe and form judgments concerning the abilities of his pupils has a superior vantage point from which to rate them on a scale of relative merit or ability. This, after all, is just what the test attempts to do. Regardless of their wishes, makers of tests are constantly forced to fall back upon this validation technique. In the long run, an educational test which consistently picks out the pupils who, in the teacher's judgment of a specific ability, are superior or inferior, probably does have significant validity.

Correlation with Ratings of Expert Judges. This procedure is related in many respects to the one discussed above. To the extent that teachers' marks are the judgments of experts, the two procedures are identical.

Correlation with Other Known Measures. This method may be utilized in fields in which extensive critical work in test development has already been done. There would be reason to doubt the validity of an achievement test in problem solving in arithmetic, algebra, or physics which did not show some relationship to achievement in problem solving as measured by other valid tests of these subjects. This is particularly true in the factual subjects. However, in certain general tool-skills, such as reading or language, it is not uncommon to find a very low correlation between two tests which are both obviously measures of certain phases of the same abilities. The explanation for this lies, undoubtedly, in the very great complexity of such abilities. There are so many different aspects of the reading or language abilities, or so many different forms in which they may be revealed, that different types of measures of the subjects may show a reasonably independent validity and yet be only indifferently related to one another. This method of test validation is most frequently used when there is available a test generally accepted as an outstandingly superior test to serve as the criterion. For example, the individual intelligence test constitutes the best basis for the validation of group intelligence tests.

An illustration of the use of correlation coefficients in the validation of a test is given in Table I for the *Powers General Science Test*. The degree to which test scores correlate with school marks and with Regents' examination marks illustrates test validation by the first and third of the statistical methods discussed immediately above. The relationships between achievement test scores and intelligence quotients show in general the degree to which pupils of high intellectual ability perform better on the test than do their less able classmates, and at least indirectly furnish validation evidence of a type discussed below—rise in percentage of success.

Correlation with Measures of Future Outcomes. This method of validation is used primarily with prognostic and sometimes with aptitude tests. As the purpose of a prognostic test is to predict future outcomes, e.g., the success of ninth-grade pupils in a course in first-year algebra, the degree

TABLE I
STATISTICAL EVIDENCE OF VALIDITY OF THE POWERS
GENERAL SCIENCE TEST⁸

<i>Form</i>	<i>Cases</i>	<i>Measure Correlated</i>	<i>Coefficient</i>
A	64	School Marks	.71 \pm .04
B	63	School Marks	.59 \pm .05
A	64	Regents' Marks	.68 \pm .05
B	63	Regents' Marks	.67 \pm .05
A	99	Intelligence Quotients	.52 \pm .05
B	97	Intelligence Quotients	.57 \pm .05
A	127	Intelligence Quotients	.47 \pm .05
B	127	Intelligence Quotients	.56 \pm .04

to which scores on the test are related to measures of the outcomes the test attempts to predict indicates the validity of the test.

Another group of validation methods primarily statistical in nature but not involving correlation coefficients is based on differences in test scores made by pupils having different subject-matter backgrounds or levels of maturity. These methods are primarily used by the maker of standardized tests.

Accomplishment of Widely Spaced Groups. One of the readily recognized evidences of validity in test content is the power of such material to reveal significant differences in the accomplishment of widely spaced groups. For example, a performance test for use in the ninth-grade wood-working shop might be validated by administering it to groups of ninth-grade pupils who have had a semester of shop work, and to similar ninth-grade pupils who have taken no industrial work in this field. If the test is valid in content, the differences in the scores made by the two groups should be significant. It is assumed, of course, that the pupils have actually learned something in the semester course in shop work. This procedure is frequently used in the validation of aptitude tests and of other tests in which rather highly specialized skills are involved.

⁸ Samuel R. Powers, *Directions for Giving Powers General Science Test*, p. 1. Bureau of Publications, Teachers College, Columbia University, New York

Rise in Percentage of Success. This procedure is based on the changes which increases in training and in maturation bring about. A valid reading test is expected to show significant increases in scores indicative of increased achievement as the tests are advanced from one school grade into the next. If second-semester algebra students do not show a higher accomplishment on a test in this field than they produced during their first semester of work in the subject, there is reason to doubt the validity of the content of the test.

Social Utility. The validation of content in terms of social utility assumes that the course of study itself is based upon that point of view. This procedure is distinctly in line with modern theory in curriculum construction. There may be occasions when the standardized test anticipates the course of study by the adoption of the social utility point of view in selection of test content, but this is unquestionably rare, and should not be expected of the test-maker.

An example of this approach to spelling test construction is the use of words which exhaustive word counts have shown to be those most widely used in written language, and therefore the words pupils need most to be able to spell correctly. Also, home mechanics tests might be based in part on the skills, such as fixing a leaking water tap, hanging a window weight, or wiring a buzzer, which activity analyses have shown to be most frequently required in the maintenance of household equipment. Again, current affairs tests should be based largely upon the events, names of persons, dates, etc., which appear most prominently in the news of the day.

Psychological and Logical Validity. There are certain subjects in which it appears to be impossible to secure an objective or statistical basis of validation. In general, these subjects are in the complex fields made up of many inter-related abilities, as language and reading. The most effective approach to the validation of the content of tests in these fields appears to be through introspective analysis. That is, a sort of arm-chair psychological dissection of the total process is made, in which as many as possible of the basic abilities are identified. In the use of this method the analysis is the initial step. Statistical refinements later make possible the se-

lection of the qualities which are best measured by objective methods.

II. RELIABILITY

A test is said to be reliable when it functions consistently. *The reliability of an examination depends upon the efficiency with which a test measures what it does measure.* This statement may appear on the surface to conflict with, or to repeat, the statement in the preceding section concerning the validity of an examination. Such is not the case, however. A test may satisfactorily test what it *does* test without to any effective degree testing what its user *attempts* to test. However, it cannot efficiently measure what it *attempts* to measure unless it efficiently measures whatever it *does* measure. This is equivalent to the statement that a test may be reliable without being valid but that it cannot be valid unless it is reliable. Therefore, *reliability is really an aspect or phase of validity.* When a reliable test is used with the type of pupils and for the purpose for which it is intended, it will also be valid. This concept is fundamentally a restatement of the fact brought out in the above section—that validity is *specific* and that it depends not only upon test content but also upon the proper use of the test. So reliability, even though it is an aspect of validity, is general, whereas validity is specific.

Reliability is most frequently expressed by the use of the coefficient of correlation. In each of the four methods presented for obtaining or estimating the reliability coefficient, it is the internal consistency or self-consistency of the test which is being evaluated. Only the general methods of obtaining the coefficients and discussions of their applications are given here. The statistical procedures involved in the use of the various coefficients are presented in Chapter XXIII.

Reliability Coefficient. The method of determining the reliability of a test is by means of correlating scores on two equivalent forms of the same test given successively by the same procedure to the same group of pupils. The resulting measure is called the *reliability coefficient* or *coefficient of re-*

liability. Thus, as is true of the validity coefficient, the reliability coefficient is simply a special application of the coefficient of correlation. It is the judgment of the authors that students interested in making a critical analysis of the reliabilities of standardized tests should do so on the basis of the correspondence of scores on two forms of the test. The resultant coefficient is likely to be safe and to be free from factors making for artificially high relationships, which sometimes result from less critical methods.

Retesting Coefficient. One method of estimating test reliability when two forms of the test are not available or cannot conveniently be given makes use of the retesting coefficient. This coefficient, which is also a special application of the coefficient of correlation, is sometimes used when only one form of a test is available. The test is given to the group of pupils twice under similar testing conditions and the retesting coefficient is the correlation coefficient between the two sets of scores. The second administration of the test should not too quickly follow the first, for a significant increase of scores may result from the previous experience with the test, but neither should it be delayed until forgetting has operated to a high degree. In any event, some increase of scores will probably result from the practice effect. Lindquist points out that this method is in general unsatisfactory, especially for achievement tests, and that it results in a spuriously high coefficient.⁴

"Chance-Half" Coefficient. A second method of estimating the reliability of a test is by means of the "chance-half" or "split-half" procedure. The test is given to a group of pupils and their scores are then obtained for two arbitrarily determined halves of the test. Usual methods of dividing a test into "chance-halves" are: (1) obtaining separate scores on the odd-numbered and on the even-numbered items, or (2) obtaining separate scores on items 1, 4, 5, 8, 9, 12, 13, etc., and on items 2, 3, 6, 7, 10, 11, etc., to equalize difficulty of the two half-scores when the items are in a scaled order of difficulty. The correlation coefficient obtained between the

⁴ E F Lindquist, *A First Course in Statistics*, pp 203-4. Houghton Mifflin Co., Boston, 1938.

two sets of scores indicates the degree of conformance between the two chance-halves of the test. The reliability coefficient which would be expected for a test as long as the two halves combined is then found by "stepping up" the correlation by means of the *Spearman-Brown Prophecy Formula*, an arbitrary formula devised for that purpose.

This method of estimating test reliability has been very popular in the past, since it involves a small amount of labor and expense. Recently, however, there has been a reaction against such a procedure. Lindquist points out that the coefficients of reliability estimated by this method are less dependable than those obtained by correlating scores on two forms of a test and are also likely to be spuriously high.⁵ Despite that fact, this is one of the most feasible methods for use with informal objective examinations for which ordinarily no second or alternate form is available.

"Footrule" Coefficient. This third method of estimating test reliability furnishes a coefficient which may in some cases be an underestimate but which is never an overestimate of the reliability coefficient. Called a "Footrule" coefficient because it admittedly is not the most accurate method, it requires the use of only three facts and measures from the test in a simple formula—the arithmetic mean and standard deviation of the scores and the number of items in the test.⁶ Because of its simplicity and because it furnishes a result of sufficient accuracy for many uses, this method is recommended for use by teachers in estimating the reliability of their informal objective examinations. The method of computing the "Footrule" coefficient is presented in Chapter XXIII.

III. ADEQUACY

Tests make no pretense of measuring every skill, ability, fact, attitude, etc., which the pupils acquire as outcomes of instruction. Such comprehensive measurement would be impossible with present measurement techniques and more-

⁵ *Ibid.* p. 203.

⁶ G F Kuder and M W Richardson, "The Theory of the Estimation of Test Reliability," (Formula 21). *Psychometrika*, 2 151-60, September 1937.

over would be relatively wasteful of time and effort. As a substitute, the same procedures of sampling as are used in many fields have been adapted to test construction. Just as a grain buyer samples a carload of wheat by taking samples from different places in the car and grading the samples in order to obtain a measure of quality for the whole carload, a test constructor measures the educational attainments of pupils by constructing test items which represent widely the types of pupil outcomes expected and accepts the scores resulting from their use as representative of the pupils' relative achievements for the entire area sampled by the test items. *Adequacy is the degree to which a test samples sufficiently widely that the resulting scores are representative of relative total performance in the areas measured.*

The diagram of Figure 3 is used to show the effect of sampling on the reliability of test exercises based on a certain limited field of information. Each of the 40 rectangular spaces in the diagram represents an item of information. Pupil A has had an opportunity to learn. The 28 shaded and the 12 unshaded rectangles represent respectively the items which he has and has not mastered. Thus, he has learned 70 percent of the facts.

It can be assumed that a very limited sampling at one end of the field is used as a test for Pupil A. If the items or exercises numbered 1 to 5 are selected, he will fail on all except item 5. However, if the last five items—16 to 20—are selected for the test, he will succeed on all except item 18. Thus, there is a variation from 20 percent to 80 percent of correct responses. Again, if he is separately tested on all the even-numbered and odd-numbered items, he will miss five and two items respectively, so his percentage scores will be 50 and 80. Finally, if he is tested on all 20 numbered items, he will miss seven and consequently have a percentage score of 65. It is to be noted that as the number of items selected is increased the pupil's success on the test more nearly approaches the actual amount of his information in this field. Thus it is clear that the extent of the sampling which the exercises in a test represent is an important factor in the accuracy of the scores. If the sampling is small, the

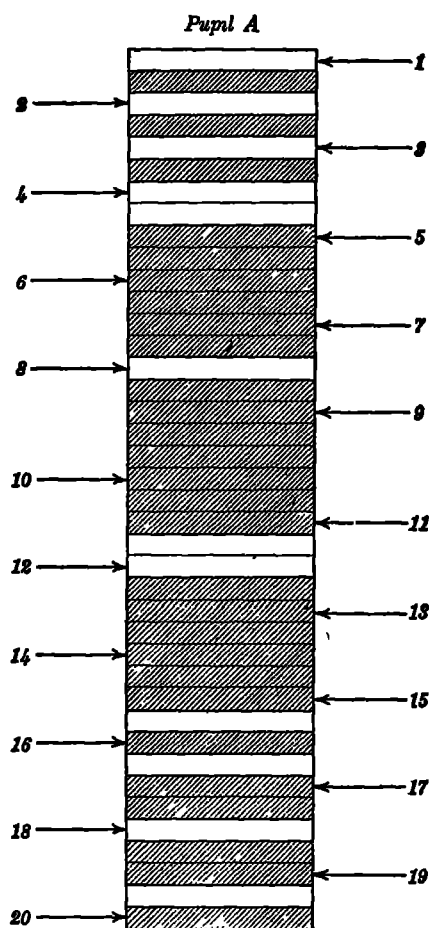


FIGURE 3 THE PRINCIPLE OF SAMPLING

scores are likely to be unfair to some pupils, if the sampling is ample, the scores are likely to be fair to all pupils.

The above illustration shows that the accuracy or consistency of test scores depends on the extent of sampling, and it was pointed out in the section on reliability that a reliable test must measure consistently. Therefore, adequacy of sampling is essential to reliability in a test, and adequacy should be considered as a phase or aspect of reliability.

IV. OBJECTIVITY

A test is objective when the teacher's personal judgment does not affect the scoring of the test. The need for the removal of the subjective factor in the marking of examination papers was recognized early in the growth of the testing movement. This recognition was one of the major factors contributing to the development of the standardized and informal objective tests. *Objectivity in a test makes for the elimination of the opinion, bias, or judgment of the person who scores it.*

In general, objective test exercises are so stated that only one answer satisfies the requirements of the statement. A common form of objective exercise is the simple recall. Other common types of objective exercises are the true-false, multiple-choice, matching and completion. The distinct advantage of selecting highly objective exercises for use in educational tests is that there can be perfect agreement as to what is the correct answer. This means that outside of purely chance errors there should be no variation in the scores assigned to a given test paper by different persons or by the same person on different occasions.

The effect of objective exercises on the accuracy of marks assigned to an examination paper is shown by an analysis of the marks assigned by a group of ten teachers to two examination papers in civics written by the same pupil over the same subject matter. One paper consisted of his answers to ten essay questions, while the other gave his answers to 40 true-false items. Each teacher marked the papers independently and at different times, so there was little chance for the mark assigned previously to carry over. The range of scores shown in Table II for the essay examination was from 76 to 90, the average of the ten scores was 83, and the average amount by which the ten scores deviated from 83 was 4.3 score points. On the other hand, one score of 30, one score of 32, and eight scores of 31 were assigned to the true-false test, so the average amount by which the scores deviated from the average score of 31 was but two-tenths of a score point. The relative objectivity of the two types of examinations is shown definitely by these findings.

TABLE II

SCORES ASSIGNED BY TEN TEACHERS TO AN ESSAY AND A
TRUE-FALSE EXAMINATION OVER THE SAME
MATERIAL IN CIVICS

Teacher	Scores Assigned	
	Essay-type (10 Questions)	True-false (40 Items)
A	84	31
B	80	32
C	88	31
D	76	30
E	82	31
F	90	31
G	85	31
H	82	31
I	83	31
J	80	31
Average	83	31
Average error (deviation from the average)	3	.2

Objectivity, as well as validity and reliability, of a test may be expressed by the use of the correlation coefficient. The coefficient obtained between scores or marks assigned to a group of papers by the same individual at two different times is sometimes called the *objectivity coefficient* or *coefficient of objectivity*. However, this coefficient is less widely used than are those for estimating validity and reliability, inasmuch as the fact is quite obvious that the best types of objective test items are relatively high in objectivity.

What was pointed out above as being true for adequacy of sampling is also true for objectivity—both are essential to

test reliability and both are therefore aspects or phases, although independent ones, of reliability.

V. ADMINISTRABILITY

The characteristics of a test which make for ease and accuracy in its administration are probably best expressed by the word administrability. While this is not one of the major criteria, it is nevertheless one which is worthy of some practical consideration.

Ease of administration must be evaluated from two distinct points of view. In the first place, the administration of the test from the standpoint of the pupil must be considered. Clear, simple, and direct instructions to the pupil should appear on the test itself, and in most cases these directions should be supplemented by other directions given orally by the examiner. If some particular type of response is called for, this should be illustrated clearly on the test by one or more samples. The nature of the responses expected of pupils should be well within their range of ability, and types of test parts requiring long and involved instructions to the pupils should in so far as possible be avoided. In the second place, the test should not ordinarily require elaborate preparatory arrangements on the part of the teacher.

If the test is standardized, the manual which accompanies it should contain specific directions for the teacher to follow verbatim in giving the test. The better standardized tests of today provide the user with carefully formulated statements on the following types of items, all of which tend to protect the pupils and the teacher against the faulty administration of the tests:

1. Number of sub-parts in the test.
2. Directions for each part of the test.
3. Fore exercises to acquaint the pupil with methods of response.
4. Directions for procedure at the bottom of each page and at the end of each test part.
5. Definite statements of time limits
6. Definite statements of total possible scores on each test part.

The mechanical features of a test frequently operate to affect definitely its ease of use in the classroom. Such fea-

tures are largely the result of the editing and printing of the test. The illustrations should be clear-cut and easily identified with the content they are supposed to amplify. Paper of good quality, preferably white bond, should be used. The page size, the length of line, and the size of type used are also mechanical features which may limit the administrability of a test.

VI. SCORABILITY

The results of a test possessing scorability should be obtainable in as simple, rapid, and routine a manner as is commensurate with their importance. It is desirable that tests be subject to accurate scoring by clerical workers or other persons not conversant with their content. Various methods of facilitating the scoring of tests, and thereby increasing their scorability, have been devised. Among these methods, discussed in Chapter VI of this volume, are the use of prepared keys, the use of separate answer sheets to be scored by hand, and the use of separate answer sheets to be scored by machine.

A convenient form of answer key or stencil should be provided for standardized tests, and the manual of directions should carry complete instructions for scoring the instrument. The scoring keys should be arranged so that easy and accurate scoring of the tests can be accomplished. Properly spaced answers on scoring keys for informal objective examinations can be prepared by filling in the correct answers on a copy of the test and converting it into a set of strip keys, cutout stencils, or a combination of the two, according to the nature of the test parts.

VII. COMPARABILITY

There are two means whereby comparability of results is established for standardized tests: (1) availability of duplicate forms of the test, and (2) availability of adequate norms. Standardized tests should be accompanied in the test manual or elsewhere by adequate tables of norms adapted in type to the age and grade levels for which the test is intended and to the types of abilities it measures. By the use of such

norms, individual pupils or class groups can be compared with average performance for pupils of similar age, of similar grade placement, or who are taking the same course. By the use of duplicate forms of a test, results from testing before and after a unit of instruction can be made comparable without the necessity of using the same test twice.

Although norms and duplicate forms are available only for standardized tests, comparability of results can be established for informal objective tests by the simple statistical procedures presented in Chapter XXIII. In a sense, a series of duplicate forms is established when different class groups are tested over a period of several years, even though the tests used from year to year may overlap considerably in content. In a sense, also, norms can be statistically established on the basis of results from any but very small classes, although such norms do not possess the reliability and wide significance of norms for standardized tests which are based upon extensive pupil populations.

The importance of comparability of test results is great, for without measures of comparability some of the major values resulting from the use of tests would be lost.

VIII. ECONOMY

Economy is certainly not one of the major criteria of a good test, but it is a factor which must be given consideration. Real economy in testing will not be achieved by indiscriminate use of cheap tests or testing methods, but it is equally true that the most costly instruments and methods are not necessarily the best. Perhaps the cost of testing should, in the long run, be computed in terms of the validity of the test per unit of cost.

There are many devices by which costs of testing can be kept low without reducing the effectiveness of a measurement program. Informal objective tests can be prepared by use of the mimeograph or gelatine plate, and some types may even be given by a blackboard method or orally. The economies of time made possible through the use of some of the scoring devices mentioned in a preceding section of this chap-

ter result in real financial saving. Cooperative testing programs operating under institutional or public educational auspices in many of the states offer testing services to the schools at cost or at very low rates. Test booklets which are not necessarily destroyed by one use are now available for certain tests, whether machine-scoring or hand-scoring is used. Therefore, an effective testing program need not depend closely upon great financial outlay.

IX. UTILITY

A test may possess adequately all of the important characteristics of a good test discussed above and yet be of relatively little value for use in a particular school situation. *A test possesses utility to the degree that it satisfactorily serves a definite need in the situation in which it is used.* Unless tests are selected or constructed for definitely conceived purposes and their results used in an intelligent attempt to bring about the desired results, they are of little value and may even, in fact, be harmful. Tests are no longer used for the mere satisfaction of curiosity and the booklets allowed to collect dust on storeroom shelves after they have been administered and scored. The modern teacher has a definite purpose in mind when he tests his pupils, and makes as effective use as possible of the results for the betterment of his pupils.

If the test is standardized, simple illustrations of the methods of interpreting the results should be given in the manual, and brief suggestions for a follow-up program of corrective instruction should also be provided when the field is one in which such remedial work is possible. Class record and tabulation sheets also add greatly to the utility of a test. If the test is one which the teacher constructs, its utility depends largely upon the foresight of the teacher in so planning the test and its use that the results will serve the needs of the local classroom.

Utility may in a sense be considered a final master criterion. It is certainly not entirely distinct from the other criteria, but it may be an effective final check on the value of the test.

TOPICS FOR DISCUSSION

1. What is meant by test validity? Define or explain validity in several ways. Is it a general or a specific concept?
2. Discuss and illustrate the two major methods by which validity is obtained in a test. What is the final or ultimate basis upon which test validation depends?
3. How does the concept of validity differ for standardized and informal objective tests?
4. Define or explain reliability as a criterion of a good test. Is it a general or a specific concept?
5. Briefly discuss the methods by which the reliability coefficient of a test can be obtained or estimated, and consider the relative merits of the several methods.
6. Is a valid test necessarily reliable? Explain. Is a reliable test necessarily valid? Explain.
7. Show how test adequacy is essential to test reliability. How is test adequacy assured?
8. Show how objectivity in a test contributes to its reliability. Characterize an objective test.
9. By what means is administrability obtained in a test? Is this an important criterion of a good examination?
10. How may scorability be obtained in a test?
11. What is meant by comparability as a criterion of a good examination? What are the two major means of attaining comparability?
12. Are norms essential to a test which possesses the characteristic of comparability? Explain.
13. What is the importance of economy as a test criterion?
14. In what way is utility in a sense the master criterion of a good test?
15. Review the criteria of a good examination and show why a good test must possess all of them if it is to serve its purpose efficiently.

SELECTED REFERENCES

- Brownell, William A., "Some Neglected Criteria for Evaluating Classroom Tests." *Appraising the Elementary School Program*. Sixteenth Yearbook of the Department of Elementary School Principals, pp 485-92. Washington, D. C. National Education Association, 1937
- Buros, Oscar Krisen (Editor), *The 1938 Mental Measurements Yearbook*. New Brunswick, N. J. Rutgers University Press, 1938.
- Buros, Oscar Krisen (Editor), *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J. The Mental Measurements Yearbook, 1941
- Engelhart, Max D., "Examinations" *Encyclopedia of Educational Research*, pp. 471-78. New York: The Macmillan Co., 1941.

- Hawkes, Herbert E, Lindquist, E F, and Mann, C R. (Editors), *The Construction and Use of Achievement Tests*, Chapter II. Boston : Houghton Mifflin Co, 1936.
- Lang, Albert R., *Modern Methods in Written Examinations*, Chapter III Boston Houghton Mifflin Co, 1930.
- Nelson, M J., *Tests and Measurements in Elementary Education*, Chapter XII New York The Cordon Co., 1939
- Odell, C W, *Educational Measurements in High School*, Chapter III. New York The Century Co, 1930
- Odell, C W, *Traditional Examinations and New-Type Tests*. New York The Century Co., 1928
- Orleans, Jacob S, *Measurement in Education*, Chapter 2 New York : Thomas Nelson and Sons, 1937.
- Ross, C. C, *Measurement in Today's Schools*, Chapter III. New York : Prentice-Hall, Inc, 1941.
- Ruch, G. M., and Stoddard, George D, *Tests and Measurements in High School Instruction*, Chapter IV. Yonkers-on-Hudson, N. Y. : World Book Co, 1927.
- Smith, B. Othanel, *Logical Aspects of Educational Measurement*. New York Columbia University Press, 1938
- Smith, Henry L, and Wright, Wendell W., *Tests and Measurements*, Chapter III. New York Silver, Burdett and Co., 1928.
- Symonds, Percival M, "Factors Influencing Test Reliability." *Journal of Educational Psychology*, 19 73-87, February 1928.
- Symonds, Percival M, *Measurement in Secondary Education*, Chapter XIV. New York : The Macmillan Co., 1928.

CHAPTER V

CONSTRUCTION OF STANDARDIZED TESTS

The following problems in the construction of standardized tests are considered in this chapter.

- a.* Meaning of the standardization process.
- b.* Securing validity of standardized test items.
- c.* Preparing equivalent forms of standardized tests.
- d.* Deriving norms for standardized tests.
- e.* Establishing final validity and reliability of tests.
- f.* Preparation of final standardized test materials.

The following treatment of the construction and refinement of standardized tests is not sufficiently detailed to afford an adequate guide to the inexperienced worker who may have ambitions to construct a standardized test. A treatment sufficiently extensive and detailed for this purpose would be a volume in itself. It should, however, be adequate to make the classroom teacher more critical of all types of measuring instruments, and at the same time a bit more appreciative of the care, the drudgery, and the expense in time and money required to produce a commercial test of a quality adequate to stand up under present-day criteria.

I. MEANING OF STANDARDIZATION

Standardization, or the process of deriving comparative norms, is commonly designated as the single important factor which distinguishes the informal from the more formal test. However, a program of standardization demands a more critical analysis of subject matter, a more careful formulation of exercise material, a more exacting refinement of the techniques of evaluating test items, more critical standards of equality of items and of test forms, and more rigid statistical analysis than are usual for the informal objective test. Thus real differences in the two types of tests appear. Such a point of view makes it clear that the mere derivation of a set of norms for a test does not in any sense

make it a standardized test. The matter of securing norms is undoubtedly the most important phase of test standardization, but it is only one of several important procedures which are closely related to the standardization process.

II. ESTABLISHING VALIDITY OF TEST CONTENT

The maker of the standardized test faces the problem of preparing an examination over content and for a group of pupils he has not specifically taught. To be reasonably certain that he is fair in the selection of items, he must take care to include only those aspects of the subject which are very likely to receive emphasis in any course. Naturally this limits at the outset the selection of test content which must be general enough to fit into any school situation in which the course is taught. As a result, the problems of the test-maker in the selection of valid content for his test are distinctly complicated.

The difficulties encountered in the selection of content for the standardized test depend to a certain extent upon the subject matter to be tested. If the field is one in which the objectives and outcomes are clean-cut and readily identified, the problem may be comparatively simple. In the event, however, that the subject is one in which the aims and outcomes are vague and general, the difficulties may be almost insurmountable, and there is little likelihood that any test based on such subject matter will have high validity. In the case of arithmetic, a subject in which the basic or fundamental facts are well known, the selection of content suitable for use in a standardized test is a relatively simple matter, and many tests of acceptable validity are available in this field. Exactly the reverse is true in certain social studies courses. The fields in which the instructional aims are specific or highly factual lend themselves readily to the construction of standardized tests. Tests in subject-matter fields of a more indefinite nature are much more difficult to validate.

In most subjects the validity of the test content is very difficult to establish by acceptable statistical or objective means. In certain fields it is practically impossible. In one

subject a certain validation procedure may be effective and acceptable; in another it may be completely unsuited for use. Makers of standardized tests have resorted to many different types of validation procedures. As these are discussed extensively in Chapter IV, they do not require further comment here.

III. CONSTRUCTING AND VALIDATING TEST ITEMS

The discussion of the problems of constructing informal objective tests in Chapter VIII points out certain principles which are to be observed in the selection of the subject-matter content for any type of objective test, whether standardized or informal objective. Therefore, methods used by the makers of standardized tests in the selection of objective item types to use for each fact, principal, relationship, etc., which they wish to test and in the actual construction of various item types are discussed only briefly here.

Test validity depends not only upon the validity of content in general but also upon the validity of the individual items of which the test consists. The validity of the items before they have been tried out with a group of pupils depends upon the ability of the test constructor to select the best objective item form for each fact or idea to be tested as well as so to construct the item that it measures the desired type of pupil behavior and has none of the weaknesses pointed out in the following pages of this chapter. Objective evidence concerning item validities, however, is secured only by the actual administration of the test in preliminary form to a large group of typical pupils and by a detailed analysis of the results. Inasmuch as many of the original items may not be satisfactory, and hence must be discarded or revised, the preliminary form of the test frequently contains many more items than will be needed in the final forms.

Objectivity. The standardized achievement test would in all probability be an impossibility were it not for the development of the objective forms of test items. Objectivity in the test exercise is such an important element in the reliability of measurement which the test affords that

it would be difficult to conceive of a standardized test made up of items not characterized by objectivity. The builder of a standardized educational test in any subject faces the very difficult problem of determining the precise form of objective technique which best fits the subject matter he wishes to test. In most cases this is a problem which can be answered only by experimentation.

After the instructional areas to be covered by the standardized test have been determined, the test-maker must proceed to break up the subject matter into elements representing the basic concepts. These important elements may then be stated in some objective form, the form selected depending to a certain degree upon the subject matter itself, and somewhat on the maturity level of the pupils with whom it is to be used. Frequently it is necessary to prepare certain of these basic concepts in two or three different objective forms, selecting for final use the types which perform best under experimental conditions. The three methods by which objectivity of test items is usually assured in the construction of standardized tests are briefly discussed below.

Uniformity of Response. Test items which appear to meet the requirements of objectivity frequently are so stated that they allow considerable variation in response. This weakness in the test items is more likely to be found in recall items than in any other form. Accordingly, in the development of recall- or completion-type items, care must be taken to formulate the statement so that it will call for a single correct response. It is usually found, however, that in spite of all possible care in this matter, the pupils will suggest many responses which are almost as exact and acceptable as any of those specified in the answer key for the test. When this happens, the keys must be changed or else the items must be further revised. Other things being equal, items which set up conditions encouraging a multiplicity of answers should be eliminated or revised.

Sparing Use of Clues and Suggestions. One of the common criticisms of the objective test item is that it contains many suggestive elements or clues which the pupils soon learn to recognize as indicating a certain type of response. Unquestionably this is a significant criticism of many objec-

tive forms. In formulating objective items, great care should be taken to see that undue suggestion of the truth or falsity of an item is not inherent in the form of the statement, although this is frequently very difficult to avoid. Some experience in the making of tests of the alternate-response types indicates that the false or negative statements are more difficult to formulate and are more likely to contain recognizable clues. The systematic use of such terms as *always*, *never*, *no*, *not*, as well as such prefixes as *un-* and *in-* in the negative forms of items may easily lead the pupil to spot them at once as clues.

Freedom from Ambiguity. The elimination of ambiguity, or the possibility of misinterpretation, is one of the most difficult problems the test-maker has to face. To keep ambiguity out of an item frequently means that he must simplify the statement. When the concept itself is simple, it is almost impossible to escape making the statement so obvious that its validity as a test item is reduced. Another aspect of the problem of ambiguity in test exercises is the fact that there are certain items which to the ignorant or poorly informed pupil are perfectly straightforward and clear but which to the critical and well-informed pupil involve implications which cloud the issue. That is, the better the pupil is informed in the field represented by the item the more likely he is to be confused by it. This is a phase of ambiguity in the statement of items which is closely tied up with item difficulty. This point is discussed below as one of the major criteria for the selection of test items.

Difficulty. The difficulty of a test item is expressed in terms of the number or percentage of pupils of a certain classification who fail to respond to it successfully. The determination of the optimum difficulty of the test items to be used in a standardized test is a problem on which there is not complete agreement among test specialists. Some test authorities prefer approximately equal numbers of items at all levels from very easy to very difficult, while others prefer to use a few easy and a few difficult items but to have the majority near the 50 percent difficulty level. They are in general agreement, however, that the test as a

whole should have about 50 percent difficulty for the average pupil.

The common practice in test construction is to attempt to prepare exercises covering a wide range of difficulty, from very easy to very difficult. Items are certainly not suitable for inclusion in the test if they are so easy that no pupil of the type on which the tests are to be used fails to respond correctly. The presence of such items would merely serve to lengthen the test without adding to the reliability of its measurement. In a similar way, items which are so difficult that no pupil is able to respond correctly should not be included in the test. Thus, items which lie at the extremes of difficulty, 100 percent failure and 100 percent success, are useless, since no one is able to tell how far beyond these limits the difficulties may lie. An item which does not in a very direct way serve to differentiate between levels of achievement has no place in an educational test, since it adds only useless dead weight.

Modern practice in the arrangement of test items tends to follow the procedure of presenting items covering a wide range of difficulty in ascending order from the very easy to the most difficult. This plan makes it possible for the lower grade or the less able pupils to respond to certain items within their level of mastery without being unduly discouraged by being confronted at the outset with exercises of prohibitive difficulty. On the other hand, it also serves to force the more able pupils to waste a certain amount of time working through a large number of items which are time-consuming but are not hard enough to bring out their real abilities. The allowance of liberal working periods for such tests tends to take care of this difficulty somewhat. Thus each pupil is allowed to work long enough to reach the level at which his abilities are taxed to the utmost. If the test items are carefully scaled in such a way that there is a gradual and continuous rise in the difficulty of the exercises, a relatively small amount of time is lost by the superior pupils in working on items which are too far below their abilities. Similarly the levels of ability of the less accomplished are revealed quite promptly and accurately. The

problems involved in the scaling or statistical evaluation of test items are rather technical and require a much more extensive treatment than can be presented in this volume.

Discriminative Power. The basic function of all measurement is to place individuals along a scale of ability or achievement in approximate accordance with their real differences in ability or achievement. Such a function implies discriminative power on the part of the test. Since tests are made up of separate items, it is clear that each item comprising a test must have this quality in a maximum degree if the total test is to possess it.

Discriminative power in a test or a test item means that a different quality or magnitude of response may be expected from individuals or groups possessing the abilities in question in varying degree. Pupils with limited ability should fail the item more often than should superior pupils. This suggests a method by which the power of a test item to discriminate or distinguish between groups of pupils may be determined. The practical implications of this procedure may be illustrated quite simply.

The illustration is based on an experimental test which has been given to a class of one hundred pupils having the normal range of ability in the subject. The tests will be corrected by the use of the answer key, and the score of each pupil in terms of the number of exercises answered correctly will be computed. On the basis of these scores, the class of one hundred pupils can be divided into two groups—those making scores above the median and those making scores below the median of the entire group. The next step involves an item count for all of the items in the test. The number and percentage of pupils in the superior group failing on Item I are determined and compared with similar data from the inferior group. A check of this type is made for all items in the test. A summary of a brief sampling of items from a typical test is given in Table III.

This table indicates that Item 1 was failed by two percent of the superior and by five percent of the inferior pupils. This item thus shows a limited power to distinguish between good and poor achievement. The fact that the item is missed by such a small proportion of all pupils indicates

TABLE III
DISCRIMINATIVE POWER OF TEST ITEMS IN PERCENTAGE
OF FAILURE BY SUPERIOR AND INFERIOR GROUPS

Item	Superior Group	Inferior Group
I	2 0	5 0
12	4 6	4 4
23	7 2	12 8
44	10 4	8.4
55	12 8	12 8
76	24 8	32 0
97	41 6	72.0
108	60 4	56 6
129	79 2	86 2
140	82 4	70 4

that its difficulty is slight. Item 44, however, is missed by a smaller percentage of poor pupils than of superior pupils. In an item of this degree of difficulty, such a failure to register the real differences in the ability of the two groups is probably serious enough to warrant the elimination, or at least the revision, of the item. Item 97 appears to be excellent, while Items 108 and 140 should certainly be subjected to further critical study.

While this method of determining the discriminative power of test items is not precisely the one used in the critical analysis of test items for standardization purposes, it may serve to illustrate the general principles involved. Incidentally, the classroom teacher who is interested in the experimental development and analysis of informal objective examinations will find in the method illustrated a very satisfactory procedure for the analysis of such test items.

IV. CONSTRUCTING EQUIVALENT FORMS

Methods of Equating Test Forms. Two or more forms of an educational test are considered to be equal or equated

when practically identical scores on each are made by the same individuals or by individuals of the same ability. This means that the forms of the test must be made up of test items which parallel one another closely in difficulty. In practice, such close equality of item difficulty in alternate forms is obtained in one of three ways.

(1) The first procedure involves the preparation of large numbers of items covering the total range of the subject matter to be tested, on the chance that there will be a sufficient number of items at each of many difficulty levels to permit of pairing items of equivalent difficulty in the alternate forms of the test. When this is done, the alternate forms of the test may be considered roughly equal in difficulty but there will be only a very general and broad equivalence of content.

(2) The second procedure involves the preparation of parallel items on certain selected, important concepts. One item may test the identification of the concept, while the other may test the identification of an additional phase of the concept or some phase of the identification of the procedure involved.

(3) A third type of procedure which permits the establishment of comparable forms of tests by the use of derived scores is mentioned here, although the complexity of the statistical techniques necessary and the variety of derived scores which are used in this way make a complete presentation impracticable at this point. It may suffice here to say that the derived scores are so established that they have constant meanings, whether or not they are obtained on the same form of the test or from the same pupil group, and that the method of establishing a "normalized group" is basic to the procedure. Several of the most widely used derived scores which are in general based on this type of procedure are presented in Chapter XXIII.

Illustration of Parallel-Item Method. An illustration of this procedure is given in the accompanying exercises from the part of the *Iowa Language Abilities Test* designed to measure the pupil's ability to decide which of two words is the correct one to use in the particular sentence.

ILLUSTRATION FROM IOWA LANGUAGE ABILITIES TEST OF PARALLEL-ITEM METHOD OF EQUATING TEST FORMS¹

FORM A	FORM B
12. (1) Whose (2) Who's book is that?	12. (1) Whose (2) Who's the boy in the dark suit?
13. (1) Leave (2) Let me go, please.	13. Please do not (1) let (2) leave me alone in this room.
16. Did you see the cat wash (1) its (2) it's face?	16. (1) Its (2) It's a dark stormy night
17. Paul went (1) into (2) in the house.	17. The man is going (1) into (2) in the post-office
19. His mother sat (1) beside (2) besides him.	19. He wanted no one near him (1) beside (2) besides his mother.
20. (1) Your (2) You're going with us next time.	20. (1) Your (2) You're helping us to do the work saved us time.

In these tests, the exercises parallel each other as to content and also quite closely as to difficulty. The items are arranged in approximately ascending order of difficulty, as represented by the number of errors made by pupils in responding to them in preliminary experimentation with the material. They are also arranged in such a manner that the two forms represent almost exactly the same difficulty as a whole, as well as almost parallel difficulty at any given point in the test. An exact equivalence of difficulty is not demanded for each pair of items, as a slight difference in difficulty for the two items of one pair may be compensated by an opposite and equivalent difference in difficulty for the items of another pair. This method of shifting and balancing the items for the two forms of the test results in a roughly scaled test of two or more forms composed of items likely to be failed by approximately the same percentages of cases. The accuracy of this method of equating test forms depends to a large degree upon the extent and the representative nature of the sampling of pupil responses used in the preliminary evaluation of the items.

¹ H. A. Greene and H. L. Ballenger, *Iowa Language Abilities Test*, Intermediate. Scheduled for immediate publication by World Book Co.

V. DERIVING TEST NORMS

Norms provide the user of a standardized test with the basis for a practical interpretation and application of the results. Unless the norms which accompany a test reflect a representative picture of the type of accomplishment to be expected, they are useless and they render the test itself useless.

Early in the history of the development of objective testing techniques, practically all that the development of a standardized test required was to give a few test exercises to a hundred or more pupils in different school systems. The results were then compiled and submitted as norms. The standardized test differed from a reasonably good informal objective test only in the fact that the former had been tried out with more pupils in a larger number of different classes. In fact, many informal examinations of the objective type meet all criteria of standard tests except that of having norms for the evaluation of their scores. Standardized tests are characterized by the fact that they are commonly accompanied by norms or tables of typical scores representative of the accomplishment which may be expected from classes similar to those used in the standardization program. However, test standardization as it is now interpreted means much more than the mere derivation of norms, although the existence of norms is still the most distinctive feature of the standardized test.

Norms are tables of information necessary for the interpretation of test scores, and are obtained by giving the particular test to a large and representative sampling of pupils in the same grades and of a type similar to the groups with which teachers will use the tests. To the extent that the sampling used in obtaining the norms was distributed over a large population in typical school situations and that the conditions under which the tests are to be administered are rigidly followed by the teachers using the tests, the norms furnish a reliable and useful basis for interpretation.

Types of Norms. The form in which the norms for a test are provided depends to a large degree upon the level in the school system at which the test is used. The norms

are also conditioned somewhat by the nature of the test itself. Tests which are designed for use in the elementary school grades are usually accompanied by two or three types of norms—age norms and grade norms, and sometimes percentile norms based on grade placement. Tests intended for use in the secondary school are more frequently provided with percentile and grade norms only. Age norms do not seem to be particularly useful at the high school and college levels, since so many factors other than age operate to affect achievement. Then too the curve of mental growth flattens out very rapidly after the fifteenth or sixteenth year, so that the increments of growth in achievement from age to age at the upper levels are relatively not significant. In place of age norms for secondary school and college tests, common practice today is to provide tables of percentile equivalents for the scores. These tables permit a very satisfactory interpretation of the test scores at these grade and age levels.

Brief discussions of grade norms, age norms, and percentile norms based on grade placement are given below. The brief illustration of how such percentile norms are derived indicates roughly the procedure used by test-makers in establishing norms for a test. Although the illustration applies to only one of the types of norms discussed, similar procedures are used in obtaining norms of the other types. As will be made evident in the following pages, percentile norms may be based on grade groups, and, though they less frequently are, on age groups, so in one sense percentile norms may become either grade norms or age norms in certain applications.

Grade Norms. In the derivation of grade norms for standard tests it is a common but not universal practice to express the norms in terms of end-of-the-year achievement. In any event, it is desirable to have the norms clearly indicate the period they are designed to cover. In the derivation of grade equivalents from the grade norms, the norms are so adjusted that the approximate progress of the pupil through the grade is indicated by the grade equivalent assigned to his score. For example, the seventh-grade end-of-the-year norm for a certain test might be 120 points, and

the eighth-grade end-of-the-year norm 140 points. A score of 130 points would be treated as equivalent to achievement halfway through the eighth grade, or 8^5 . A score of 140 would be recorded as 8^{10} . The exponent in each case represents the number of tenths of a grade of achievement revealed by the score in terms of the grade norms.

The grade norms established for most of the commonly used achievement tests are based on the median scores obtained by giving the tests to large groups of pupils in each grade. Such norms provide the basis for the interpretation of class scores as well as of individual accomplishment.

Age Norms. The problem of establishing age norms for tests has caused a great deal of difficulty. An early practice in the preparation of age norms involved the regrouping of all pupils used in the grade tabulation into chronological age groups regardless of grades. The test scores of these chronological age groups were then tabulated, and the means or medians computed. These results were then used as the basis for setting up tables of the scores corresponding to the several age groups. It soon became apparent, however, that many factors other than age were operating to influence the average achievement of pupils grouped in grades. Such factors as over-ageness, retardation, and a lack of balance between retardation and acceleration were all present. It was found, for example, that while the average chronological age of a seventh-grade pupil might be 13 years and 6 months at the end of the school year, the average test score of pupils of 13 years and 6 months was not at all the same as the end-of-the-year score for the seventh grade. This caused much confusion in the interpretation of grade and age data. Obviously some method of taking care of this problem was required.

Crawford has shown that test norms based on unselected cases are definitely affected by the relative balance of under-age and over-age pupils.² The actual achievement of the under-age pupils is significantly superior to that of over-age groups in a given grade, as might be expected. For the makers of standardized tests the useful implication from this

² John R. Crawford, *Age and Progress Factors in Test Norms*. University of Iowa Studies in Education, Vol. IX, No. 4, University of Iowa, Iowa City, 1934.

study lies in the fact that it makes very apparent the real need for differential norms, or norms which will take into account wide differences in mental ability or school progress within the grade.

An example of norms given in terms of both grade and age is shown in Table IV for the *Metropolitan Arithmetic Tests*. It is possible to determine quickly from such a table both the grade equivalent and the age equivalent for any given score.

TABLE IV
AGE AND GRADE EQUIVALENTS FOR PART OR SUB-TEST SCORES
ON THE METROPOLITAN ARITHMETIC TESTS³

Part or Sub- test Score	Grade Equiv- alent	Age Equiv- alent	Part or Sub- test Score	Grade Equiv- alent	Age Equiv- alent	Part or Sub- test Score	Grade Equiv- alent	Age Equiv- alent	Part or Sub- test Score	Grade Equiv- alent	Age Equiv- alent
11	3 1	8-4	34	5 4	10-11	57	7 7	13-2	80	10 0	15-3
12	3 2	8-6	35	5 5	11-0	58	7 8	13-4	81	10 1	15-4
13	3 3	8-7	36	5 6	11-1	59	7 9	13-5	82	10 2	15-5
14	3 4	8-8	37	5 7	11-3	60	8 0	13-6	83	10 3	15-6
15	3 5	8-9	38	5 8	11-4	61	8 1	13-7	84	10 4	15-7
16	3 6	8-11	39	5 9	11-5	62	8 2	13-8	85	10 5	15-8
17	3 7	9-0	40	6 0	11-7	63	8 3	13-9	86	10 6	15-9
18	3 8	9-1	41	6 1	11-8	64	8 4	13-10	87	10 7	15-10
19	3 9	9-3	42	6 2	11-9	65	8 5	13-11	88	10 8	15-11
20	4 0	9-4*	43	6 3	11-10	66	8 6	14-0	89	10 9	16-0
21	4 1	9-5	44	6 4	12-0	67	8 7	14-1	90	11 0	16-1
22	4 2	9-7	45	6 5	12-1	68	8 8	14-3	91	11 1	16-2
23	4 3	9-8	46	6 6	12-2	69	8 9	14-4	92	11 2	16-3
24	4 4	9-9	47	6 7	12-3	70	9 0	14-5*	93	11 3	16-4
25	4 5	9-11	48	6 8	12-4	71	9 1	14-6	94	11 4	16-5
26	4 6	10-0	49	6 9	12-6	72	9 2	14-7	95	11 5	16-6
27	4 7	10-1	50	7 0	12-7	73	9 3	14-8	96	11 6	16-7
28	4 8	10-3	51	7 1	12-8	74	9 4	14-9	97	11 7	16-8
29	4 9	10-4	52	7 2	12-9	75	9 5	14-10	98	11 8	16-10
30	5 0	10-5	53	7 3	12-10	76	9 6	14-11	99	11 9	16-11
31	5 1	10-7	54	7 4	12-11	77	9 7	15-0	100	12 0	17-0
32	5 2	10-8	55	7 5	13-0	78	9 8	15-1			
33	5 3	10-9	56	7 6	13-1	79	9 9	15-2			
*Values below 9-4 and above 14-5 are extrapolated											

³ Richard D. Allen, et al., *Directions for Administering Metropolitan Achievement Tests, Intermediate and Advanced Arithmetic*, p. 5. World Book Co., Yonkers-on-Hudson, N. Y., 1933.

Percentile Norms. Percentile norm tables show for a wide sampling of pupils in a certain grade of school or who are taking a certain high school course either (1) the percentage of pupils exceeding each score or each of a number

TABLE V
RATE PERCENTILE SCORES FOR SCHRAMMEL-GRAY
HIGH SCHOOL AND COLLEGE READING TEST⁴
RATE PERCENTILE SCORES

Grades	7	8	9	10	11	12	College Freshmen
Percentiles							
99	185	185	185	185	185	185	185
95	183	183	184	184	184	184	184
90	182	182	182	182	182	182	183
85	180	180	181	181	181	181	182
80	172	173	175	177	180	180	181
75	163	164	166	169	175	175	180
70	156	157	158	163	170	170	177
65	151	152	154	158	163	165	173
60	145	146	150	154	158	159	169
55	138	140	145	150	154	156	162
50	133	135	139	144	150	153	158
45	128	130	135	139	146	149	154
40	123	124	131	135	141	143	152
35	116	117	126	131	136	138	147
30	111	112	121	125	131	133	142
25	107	107	116	120	125	128	137
20	101	102	110	114	118	122	130
15	94	94	104	107	112	114	123
10	86	86	97	100	106	106	114
5	75	75	89	92	95	97	105
1	55	56	70	73	80	82	92
No. Cases	252	256	1426	1016	728	715	1424
S. D.	41 4	42 2	37 0	36 3	37 0	34 8	31 8
Read table thus: In grade 7 a score of 185 merits a 99th percentile rank, one of 183, a 95th percentile rank; one of 182, a 90th percentile rank, one of 133, a 50th percentile rank, and so on.							

⁴ H E Schrammel and W H Gray, *Manual of Directions Schrammel-Gray High School and College Reading Test* Public School Publishing Co, Bloomington, Ill, 1940.

of equally-spaced scores, or (2) the score below which certain percentages of pupils fall, as 10 percent, 20 percent, etc. Although percentile norms are customarily presented in one or the other of these methods, there is great variety in the actual form of such tables. Table V presents an illustration of percentile norms of the second type.

Illustration of the Derivation of Norms. The derivation of percentile norms is initiated by the simple process of

TABLE VI

DISTRIBUTIONS OF SCORES ON THE IOWA GRAMMAR INFORMATION TEST⁵

Mid-Point of Class-Intervals	Grades			
	9	10	11	12
75			1	1
72			2	4
69		1	5	6
66	1	2	6	4
63	3	2	7	8
60	3	5	10	12
57	7	8	11	13
54	10	12	10	19
51	14	7	14	14
48	13	14	19	23
45	15	12	27	17
42	18	25	26	16
39	23	20	22	17
36	27	17	17	10
33	25	18	10	14
30	28	19	12	12
27	24	18	10	10
24	21	15	12	8
21	16	10	9	2
18	17	8	3	4
15	12	11	1	2
12	11	8	1	
9	5	2		
Total	293	234	235	216

⁵ Fred D. Cram and H. A. Greene, *Iowa Grammar Information Test*. Published by Bureau of Educational Research and Service, University of Iowa

Percentile Norms. Percentile norm tables show for a wide sampling of pupils in a certain grade of school or who are taking a certain high school course either (1) the percentage of pupils exceeding each score or each of a number

TABLE V
RATE PERCENTILE SCORES FOR SCHRAMMEL-GRAY
HIGH SCHOOL AND COLLEGE READING TEST ⁴
RATE PERCENTILE SCORES

Grades	7	8	9	10	11	12	College Freshmen
Percentiles							
99	185	185	185	185	185	185	185
95	183	183	184	184	184	184	184
90	182	182	182	182	182	182	183
85	180	180	181	181	181	181	182
80	172	173	175	177	180	180	181
75	163	164	166	169	175	175	180
70	156	157	158	163	170	170	177
65	151	152	154	158	163	165	173
60	145	146	150	154	158	159	169
55	138	140	145	150	154	156	162
50	133	135	139	144	150	153	158
45	128	130	135	139	146	149	154
40	123	124	131	135	141	143	152
35	116	117	126	131	136	138	147
30	111	112	121	125	131	133	142
25	107	107	116	120	125	128	137
20	101	102	110	114	118	122	130
15	94	94	104	107	112	114	123
10	86	86	97	100	106	106	114
5	75	75	89	92	95	97	105
1	55	56	70	73	80	82	92
No Cases	252	256	1426	1016	728	715	1424
S. D.	41 4	42 2	37 0	36 3	37 0	34 8	31 8
Read table thus In grade 7 a score of 185 merits a 99th percentile rank, one of 183, a 95th percentile rank; one of 182, a 90th percentile rank, one of 133, a 50th percentile rank, and so on.							

⁴ H E Schrammel and W H. Gray, *Manual of Directions Schrammel-Gray High School and College Reading Test*. Public School Publishing Co, Bloomington, Ill, 1940

of equally-spaced scores, or (2) the score below which certain percentages of pupils fall, as 10 percent, 20 percent, etc. Although percentile norms are customarily presented in one or the other of these methods, there is great variety in the actual form of such tables. Table V presents an illustration of percentile norms of the second type.

Illustration of the Derivation of Norms. The derivation of percentile norms is initiated by the simple process of

TABLE VI

DISTRIBUTIONS OF SCORES ON THE IOWA GRAMMAR INFORMATION TEST⁵

Mid-Point of Class-Intervals	Grades			
	9	10	11	12
75			1	1
72			2	4
69		1	5	6
66	1	2	6	4
63	3	2	7	8
60	3	5	10	12
57	7	8	11	13
54	10	12	10	19
51	14	7	14	14
48	13	14	19	23
45	15	12	27	17
42	18	25	26	16
39	23	20	22	17
36	27	17	17	10
33	25	18	10	14
30	28	19	12	12
27	24	18	10	10
24	21	15	12	8
21	16	10	9	2
18	17	8	3	4
15	12	11	1	2
12	11	8	1	
9	5	2		
Total	293	234	235	216

⁵ Fred D. Cram and H. A. Greene, *Iowa Grammar Information Test*. Published by Bureau of Educational Research and Service, University of Iowa.

giving the test to be standardized to a large and unselected group of pupils in the grades for which the norms are desired and under the conditions which are to operate in the later use of the tests. Table VI gives frequency distributions of the scores made on the *Iowa Grammar Information Test*, by a sampling of an unselected high school population.

TABLE VII
PERCENTILE NORMS
(Based on Data of Table VI)

Percentiles	Grades			
	9	10	11	12
99	63	66	70	73
90	51	55	60	63
80	45	48	53	58
75	42	46	50	56
70	40	43	48	54
60	36	40	45	49
50	33	36	43	46
40	30	32	40	42
30	26	28	36	39
25	24	26	33	36
20	22	24	30	33
10	16	19	24	27
1	9	12	15	18

When these distributions were treated statistically, the percentile norms shown in Table VII resulted. These can be called "skeleton" norms, because of the fact that gaps appear in both the percentile and the score columns of the table. The norms are arranged to show not only median achievement for each grade but also each decile of achievement, i.e., 10th, 20th, 30th, etc. percentiles. Figure 4 shows in graphic form the median, 25th and 75th percentile points in terms of scores made by the sampling of pupils on this test.

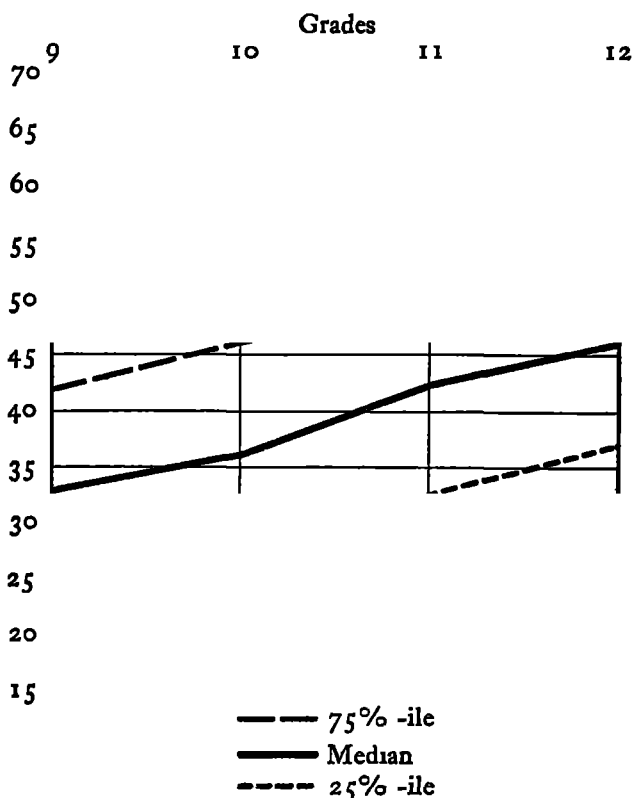


FIGURE 4. GRAPHIC REPRESENTATION OF THE MEDIAN, 75TH AND 25TH PERCENTILES GIVEN IN TABLE VII

This brief illustration of how percentile norms are obtained represents in general the method used in deriving grade norms and age norms, although the statistical procedures vary according to the type of norms derived.

Norms vs. Standards. The use of the term *standardized* in the discussion of tests of the type for which norms are provided has led to the development of a careless tendency to use the words "standards" and "norms" as synonyms. The process of securing the data for the critical analysis of tests and the derivation of suitable norms is properly known as *standardizing*. However, the term "standard," when used to refer to a level of pupil achievement, implies an

ultimate goal to be achieved. Standards may not actually be reached by any individual, but they are levels of achievement toward which to strive. *Norms are the levels of achievement which typical pupils actually attain.* When considered in the light of these definitions, it is clear that there are few tests which are accompanied by standards. It might be more nearly the truth to call the process of securing these comparative scores known as "norms" by the more descriptive name of "normalizing."

Possibly one of the best illustrations of the differences between standards and norms is to be found in the field of arithmetical computations. The standard of arithmetical accuracy is naturally 100 percent, for most such computations containing error are useless. However, the actual norm of arithmetical accuracy of computation on a well-known test is from 65 to 70 percent for the junior high school grades. That is to say, pupils of these grades work these particular kinds of arithmetical examples with an accuracy of from 65 to 70 percent instead of the desired ultimate goal of 100 percent accuracy.

It should be recognized that a norm does not necessarily represent a satisfactory level of achievement. This is particularly true of those larger schools in which instruction and classroom environments are superior and in which pupils, largely because of their satisfactory home environments and heredity, have superior abilities. In any event, teachers should encourage pupils to make the most of their abilities and to surpass test norms whenever they can. Even when a class has average performance which is just at the norm on a test, representing only the attainment of what is expected from a typical class, approximately half of the pupils will still be below the norm of achievement.

Standards are of two general types. In the first place, there are certain standards of achievement, or minimum essentials, which have been fairly generally accepted by school people for such abilities as handwriting and, in less objective form, reading, spelling, and arithmetic. Although these standards are usually based on the results of standardized testing, and may make use of norm tables for their establishment, they frequently are conceived of as represent-

ing the minimum quality and perhaps speed of performance which will adequately equip the pupil for post-school life. For example, the widely accepted standard in handwriting is a quality of 60 on the Ayres' *Scale for Measuring the Handwriting of School Children* at the rate of 70 letters per minute. Although the quality of 60 at the given rate on this scale is approximately the norm for pupils completing the sixth grade, it is also thought of as the standard or minimum ability which should be attained by all pupils before they finish school.

In the second place, the standard in any school subject or form of pupil achievement may be a definitely formulated, although probably subjective, or even only a vaguely conceived, idea in the mind of the teacher or principal concerning his expectations of his pupils. In this sense, standards are extremely variable and differ from school to school, teacher to teacher, and even, as his ideas and pupil groups change, from year to year for the same teacher.

The modern emphasis upon providing for each child as an individual the type of instruction best adapted to his abilities, interests, present and future needs, etc., rather than upon the molding of all pupils into the same achievement pattern, has reduced the reliance of school people upon standards. The attempt is rather to furnish maximum aid to each child in the development of his potentialities and to evaluate his achievement in terms of himself as an individual.

VI. ESTABLISHING FINAL VALIDITY AND RELIABILITY

The procedures discussed above, although sometimes complex and always time consuming, are prerequisite to the final steps in publication of a standardized test. After these steps have been carried out, the final forms of the test given to a representative group of pupils, and the norms derived on the basis of their scores, it remains for the test-maker to obtain final evidence concerning the validity and reliability of the test and the reliability of the norms. Although careful and accurate work on the preliminary steps should make reasonably certain that these important criteria will be satisfied, it is nevertheless essential that these steps be per-

formed as a final check up and that their results be reported to users and prospective users of the test to enable them better to evaluate it.

The interlocking and complex nature of some of these final steps makes necessary only a brief presentation here of the most important aspects of this final check up. Chapter XXIII presents a more comprehensive treatment of the methods of determining the reliability of tests and of norms.

Validity of the Test. If the test is one for which validity coefficients of one or more of the types discussed in Chapter IV will be meaningful, such validity coefficients should be obtained. In some instances this might require the administration of some other test to the group of pupils on which the test is standardized, and the comparison of results obtained. In other cases it may require a comparison of test scores with course marks or teachers' ratings of pupils. Evidence concerning validity is also found in test norms which consistently show higher average scores with advancement in age and grade placement of the standardization group of pupils. In any event, evidence must be obtained directly or indirectly to show that the test measures what it purports to measure.

Reliability of the Test. Test reliability must be established for the final form or forms of the instrument. Techniques such as those presented in Chapter XXIII and even some of the more refined methods might be used. The reliability of measurement, which gives an indication of the accuracy of the scores obtained on the test, should also be determined. The purpose of such procedures is to establish the fact that the test measures accurately and consistently.

Reliability of the Norms. One of the major problems in the derivation of norms for standardized tests is the matter of the reliability of the norms themselves. Possibly the statement of this problem would be made clearer if the word *universality* were substituted for reliability in the foregoing statement. Reliability implies consistency, but universality reflects the generally representative nature of the norms. An otherwise excellently made test may be limited in its usefulness through the fact that the norms are not sufficiently representative. It may be that it is hopeless to expect to

produce norms which are so generalized that they represent suitable bases of comparisons wherever they may be found or used. However, the only hope lies in one of two directions. One is to sample so widely in the possible areas of population likely to use the test that practically every type and character of pupil and school situation is included. The other is to recognize the practical difficulties in the way of making a general norm fit all types of situations, and to select the population used in the derivation of the norms to represent deliberately chosen types of school situations. It would be impractical to expect pupils from small school systems with little or no laboratory or shop equipment to achieve at a level comparable to that expected of pupils from schools with large, well-equipped laboratories. The solution of this problem may lie in the establishment of representative norms for different types of courses and schools.

A further practical problem faced by the test-maker is the matter of determining how large a population must be sampled before he has a basis for reliable norms. For this purpose, the method of "cumulative sampling" may be used. In this procedure additional samplings are added to the distribution until a place is reached where the means (averages) cease to vary when more cases are added. Naturally a norm based on one case would be unreliable. The addition of one more case would help but the result would still lack reliability. A point is finally reached, however, where the addition of more cases does not make any significant change in the value of the consecutive arithmetic means, and it is at that point that reliability of the norms is attained.

VII. PREPARATION OF FINAL TEST MATERIALS •

Preparation and Printing of the Test. The final steps of preparing and printing the test are frequently performed prior to the testing of the groups of pupils upon which the norms are based. Whether that is the case or whether the test is given to the standardization group in an experimental form, the test must be very carefully prepared and printed or otherwise reproduced. Attention to the test format, consistency and adequacy of directions to the pupils, of timing,

of item numbering, and many other details is extremely important.

Preparation and Printing of Accompanying Materials. Such materials as a manual of directions, scoring keys, a class record form, and sometimes answer sheets commonly accompany standardized tests. These materials must be consistent with the test itself if the total job is to be well done. Unfortunately, publishers of standardized tests are not uniformly successful in this last but very important step, so the result sometimes is that the various materials for a test are not properly integrated with the test itself.

The manual of directions usually contains evidence concerning the validity and reliability of the test, directions to the tester for administering the test and scoring the results, tables of norms for use in interpreting the results, and frequently explanations concerning the major values and uses of the tests. It is apparent, then, that a carefully prepared manual, or the presentation of these materials in some other form, is of great importance. As the tendency on the part of standardized test users is increasingly toward placing the burden of proof concerning the test upon its author and publisher, and as these expository materials are found primarily in the manual and other accompanying materials, the extreme importance of this final step of test standardization is clearly evident.

TOPICS FOR DISCUSSION

1. What basic differences distinguish the informal objective examination from the standardized test?
2. Show how the process of standardization involves much more than the mere establishment of norms for a test
3. Indicate why the validation of content for standardized tests is more difficult for some school subjects than for others
4. Discuss three major characteristics of test items which possess objectivity.
5. What is desirable with respect to the difficulty of items in a standardized test?
6. Show how discriminative power in a test item contributes to its validity.
7. What reasons can you suggest for the preparation of several equivalent and interchangeable forms of a standardized test?
8. How are equivalent forms of standardized tests prepared?

9. What is the nature and importance of standardized test norms?
10. Discuss the three major types of test norms and illustrate each.
11. What factors appear to determine the type of norms which should be supplied with a standardized test?
12. Distinguish clearly between norms and standards of achievement. How should they be used?
13. What is the importance of determining the validity and reliability of standardized tests in their final forms?
14. Discuss methods by which the reliability of norms for standardized tests can be assured or made reasonably certain
15. Discuss the preparation of standardized tests and accompanying materials in their final form.

SELECTED REFERENCES

- Conrad, Herbert S, "Norms" *Encyclopedia of Educational Research*, pp 773-80 New York The Macmillan Co, 1941
- Judd, Charles H, *Educational Psychology*, Chapter 28. Boston Houghton Mifflin Co, 1939
- Lang, Albert R, *Modern Methods in Written Examinations*, Chapter XII. Boston Houghton Mifflin Co, 1930
- Lee, J Murray, *A Guide to Measurement in Secondary Schools*, Chapters XII-XIII New York D Appleton-Century Co., Inc, 1936.
- McCall, William A, *Measurement*, Book II. New York The Macmillan Co, 1939.
- Odell, C. W, *Traditional Examinations and New-Type Tests*, Chapter III. New York The Century Co, 1928.
- Ross, C. C., *Measurement in Today's Schools*, Chapters IV, X. New York. Prentice-Hall, Inc., 1941.
- Ruch, G. M, "Minimum Essentials in Reporting Data on Standard Tests." *Journal of Educational Research*, 12 349-58, December 1925.
- Ruch, G M, *The Objective or New-Type Examination*, Chapter VII. Chicago Scott, Foresman and Co, 1929
- Ruch, G M, and Stoddard, George D, *Tests and Measurements in High School Instruction*, Part IV. Yonkers-on-Hudson, N. Y.. World Book Co, 1927
- Russell, Charles, *Standard Tests*, Chapter V. Boston: Ginn and Co., 1930
- Tiegs, Ernest W., *Test and Measurements for Teachers*, Chapter V. Boston Houghton Mifflin Co., 1931
- Wood, Ben D, "The Need for Comparable Measurements in Individualizing Education." *Educational Record*, 20 14-31, Supplement No. 12, January 1939.

CHAPTER VI

USING STANDARDIZED TESTS IN THE CLASSROOM

This chapter deals with the following points concerning the classroom uses of standardized tests

- a.* Tests in relation to classroom instruction.
- b.* Instructional uses of achievement tests.
- c.* Planning the testing program.
- d.* Selecting the tests.
- e.* Administering the tests.
- f.* Scoring the tests.
- g.* Analyzing and interpreting test results.

The value of the educational test to the classroom teacher is directly proportional to the extent to which the results obtained are translated into improved teaching practices by the teacher and improved learning conditions for the pupil. Problems of securing these results in the most effective and economical manner are treated in this chapter primarily in terms of the contribution of the standardized test as an instructional instrument.

Tests in Relation to Classroom Instruction. The value of standardized educational tests for administrative purposes, as supervisory instruments, as aids to school surveys, and as research equipment has been emphasized so generally that very often the classroom teacher overlooks their real value in the solution of his own instructional problems. Yet this is where the most vital and important uses of such tests are to be found. The development of reliable, valid, and highly detailed measuring instruments has caused the teacher to modify his previous conceptions of the uses of standard tests. Earlier experience with the more formal types of educational tests sometimes led the teacher to feel that tests were merely time-consuming devices used for checking up on his teaching efficiency, and from which he at no time received any constructive help in the improvement of his instruction. The more modern conception of standard tests is quite in contrast with this idea. It implies their continuous use as in-

struction progresses. In a sense this means a continuous testing program, for experience with the other conception of the use of tests indicates that to be the only way by which standard tests will ever come to function at their highest efficiency as instructional instruments in the classroom.

I. INSTRUCTIONAL USES OF ACHIEVEMENT TESTS

For Pupil Guidance. Schools are under constant criticism for their apparent failure to identify the special abilities of their pupils and to challenge these children to greater efforts. This is one aspect of educational guidance. It is also charged that little or no attempt is made to direct children away from fields in which they apparently have little aptitude. With the modern objective devices now available for the measurement of general as well as specific abilities of children, neither of the situations needs to exist. Teachers, principals, and administrators have found that test records obtained early in the institution's contact with the student prove to be extremely valuable aids in handling disciplinary cases and in helping students to adjust themselves in many other ways. Most administrative problems arise through the failure of the school system properly to stimulate and occupy the pupil's mind. Many of the reasons for such difficulties may be made clear to the teacher by the wise use of properly selected tests. The necessary adjustments can then be made to correct a situation which need not exist if properly handled.

For Individual Pupil Diagnosis. Closely connected with the use of the test for pupil guidance is its use for the determination of the difficulties and variabilities of each individual pupil. While in general individual differences are not so marked as to preclude reasonably efficient class instruction, the more that is known about each child's weaknesses and strengths the greater are the possibilities for success on the part of the teacher instructing the group. The test results should be studied especially in the light of each pupil's individual attainments and points of difficulty. The critical analysis of each pupil's test score may very likely be a means of clearing up wholly unsuspected troubles which would otherwise continue to hamper the child and to reduce his

chances for proper advancement. Although this type of individual diagnosis has great possibilities, it becomes valuable only when it is definitely tied up with remedial material so devised that each child may be aided in correcting his own weaknesses.

Examples of two diagnostic profile charts are given in the accompanying illustrations of interpretative materials provided with two achievement tests of the survey type. Both profile charts serve analytic or general diagnostic rather than specific diagnostic functions, but it should be remembered that survey tests can be diagnostic only in the broad sense discussed in Chapter II. The profile chart for the *Progressive Reading Test* furnishes places for recording graphically evidence of a pupil's grade placement on total reading, on the two sub-total measures of vocabulary and comprehension, and on seven part scores. The chart for the *Gray-Votaw General Achievement Tests* furnishes positions for the graphic recording of the pupil's grade placement on total achievement and on achievement in six separate major areas of elementary school subject matter.

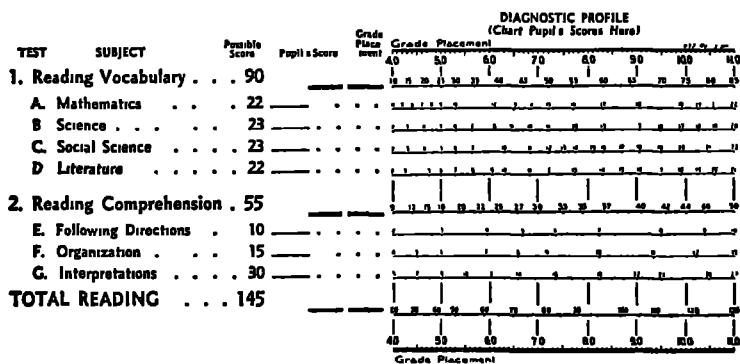


FIGURE 5 DIAGNOSTIC PROFILE CHART FOR PROGRESSIVE READING TEST¹

For Pupil Gradation. Teachers and supervisors find the problem of pupil placement one of the most difficult situations which they have to face. The indefinite lines of division between the grades and the wide overlapping of ability

¹ Ernest W. Tiegs and Willis W. Clark, *Progressive Reading Tests*, Intermediate. Published by California Test Bureau, 1937.

within the grades reveal that our technique of pupil classification has been extremely crude. Practically all investigations have shown the typical grade assignment of pupils to be inaccurate and unsatisfactory. This could scarcely be otherwise in view of the methods commonly used. Chronological age and physical size are the two pupil characteristics used most frequently even under conditions where much more comprehensive information might be readily secured. The proper grade placement of pupils implies that in so far as possible individuals who are normal for their group will be placed together. This means that pupils who are approximately alike in their chronological age, their educational achievement, and their physiological, mental, social, and moral development should ordinarily be placed together for instructional purposes. Not all of these qualities lend themselves readily to objective measurement, but a number of them do, and within these limits the results of objective measurements should be used in determining the pupil's placement in his group. Through the development of reliable grade and age norms, based upon the achievement of groups of children on standardized tests, a valuable instrument for the establishment of much more exact grade lines is made available.

By means of a simple procedure, results from several standardized tests, even though expressed in different units, can be combined into a simple graphic record and used as a valuable aid in pupil classification. The technique is simple and valuable, regardless of whether a complete reclassification of the school or grade is planned, or merely the proper placement of a few new pupils entering the system for the first time. Means for combining test scores in a graphic record are discussed in Chapter XXIII.

For Class Analysis and Diagnosis. Very often a teacher, at the beginning of a school term, wishes to obtain advance information concerning the proficiency of his classes in certain subjects and their general preparation for the work. It is essential for him to know their weaknesses and their strengths in order so to direct their work that the best results will be obtained. He needs to know the background his pupils have been given for the work they will be expected to master

during the ensuing year. Many standard tests are now available which permit this type of use. It is not always necessary for a teacher to employ a special diagnostic test to secure the required general data on the relative abilities of the class. For example, the ability to read silently is the basis of proficiency in so many subjects that he should certainly secure a picture of the reading ability of the class. The results should indicate whether the class as a whole or the individual members of the class are able to interpret the printed page with facility, and so carry on their work without great assistance. It is also possible to test in a like manner for other general qualities, as well as for knowledge of specific subject matter.

Not only is this preliminary diagnosis of great value to the teacher, but it has also been found desirable and valuable to check progress or advancement from time to time by means of objective tests. Tests of achievement will reveal whether the class as a group is moving together, or whether there are more or less well-defined sub-groups which seem to need special attention. Frequently such conditions furnish a justifiable basis for dividing a grade or class into sections for such corrective treatment. Where classes are divided into several sections, as is often done in larger schools, many competent educators feel that the pupils should be arranged so that groups of approximately equal ability are placed together. The objective test is the best means for making this adjustment so that pupils can move forward in groups of nearly equal proficiency.

An illustration of this need is given in Table VIII, which shows the rather startling range of ability found in a typical ninth-grade class for results from the *Terman Group Tests of Mental Ability*. A teacher confronted with a class ranging in mental ability from above twelfth grade to below seventh grade is faced by a hopeless task if he is expected to bring all members of this class up to the same level of proficiency. Particularly is this true when the range of ability is wholly unsuspected or measured, as it is in so many cases, by guess rather than by reliable tests. The systematic use of tests for class diagnosis constitutes a real source of professional protection to the classroom teacher.

TABLE VIII
DISTRIBUTION OF INTELLIGENCE IN A NINTH-GRADE CLASS
IN TERMS OF AVERAGE GRADE PLACEMENT

Grade Location	Number of Pupils
Above 12	6
12	4
11	12
10	8
9	19
8	8
7	6
Below 7	4
Total Number of Cases	67

For Group Comparisons. Since the earliest beginnings of group instruction, classroom teachers have wished to know just how their pupils have compared in attainment with other similar pupils and classes. Until standard tests were developed, it was practically impossible to secure this information. Now the giving of a standard test in arithmetic, spelling, reading, or other school subjects makes fairly easy a comparison of the results from a class with the norms established for the subject and grade.

Comparisons with other classes within the system in which the teacher is working, within the same building, and even between different sections of classes in charge of the same or different teachers can be made on a basis of objective norms that have been derived for the various tests. Another sort of comparison which is even more useful is that between the attainment of a class at the beginning and the end of a semester's or a year's work, or at shorter intervals in the course of a semester. Each of these comparisons has its own peculiar value in assisting the teacher to determine the relative attainment and progress of his class at a given time.

For Measuring the Efficiency of Learning. Such general comparisons as are cited above are of great value in themselves, but equally important is the determination of ways and means by which the act of teaching itself may be improved. Ambitious teachers everywhere are looking for

the best methods of instruction in their fields. Teaching methods, which in the last analysis must be studied only in the classroom by the classroom teacher, can be effectively evaluated by means of standardized tests. Instructional units within the course of study should also be evaluated. The measurement of the effect of certain types of drill exercises and the determination of the specific strengths or weaknesses of groups or classes constitute uncounted opportunities for the use of these valuable instructional devices. Only through the use of large quantities of such available material in connection with such simple yet significant investigations by the classroom teacher will the multitude of now unanswered questions be answered.

II. PLANNING THE TESTING PROGRAM

Steps in a Testing Program. A tentative outline for a testing program is presented here as a suggestion for the general organization of the work : .

1. Determine how, and what types of, test data will be valuable in the solution of instructional or classroom problems which have arisen.
2. Select the best available tests for the purpose
3. Make careful preparation, and then administer the tests.
4. Score the tests.
5. Tabulate the scores, and analyze and interpret the results
6. Use the results and interpretations in the elimination or improvement of the conditions revealed.

Clear-Cut Teaching Problem as Basis. One of the most common errors made by teachers and supervisors is the inauguration of a testing program without the formulation of a clear-cut problem, the solution of which can be most advantageously reached through the use of standard tests. The problem should be clearly defined, for the testing program will thus be more limited in extent and more intensive. The best constructive supervisory work will result from careful intensive cultivation of a limited field. If the work is undertaken in this way, much time will be saved and one of the most common criticisms—that the time of the pupil and teacher is taken for the testing and nothing ever comes of it—will be avoided. Both teacher and pupils have a right to

profit from a knowledge of the conditions revealed by the testing.

Illustrations of Problems. Problems suitable to form the basis for a testing program are to be found in almost all of the fields of education. Frequently these problems overlap. It is not uncommon for one test or a series of closely allied tests to contribute to the solution of several problems. The problems listed below are classified in accordance with their interest to administrators, supervisors, and teachers. It is clear, of course, that the list is not exhaustive.

**A. PROBLEMS PRIMARILY OF INTEREST TO TEACHERS
AND SUPERVISORS**

1. The discovery and diagnosis of defects of individual pupils in the various subjects or in particular phases of a subject as the basis for a remedial program.
2. The determination of how the pupils and the class compare with the norms in the different subjects.
3. The determination of the progress of the class in the different school subjects over a given period.
4. The determination of whether different phases of subjects are being properly or unduly stressed, as indicated by relative accomplishment of the pupils.
5. The determination of whether the pupils are working to capacity.

**B. PROBLEMS PRIMARILY OF INTEREST TO ADMINISTRATIVE
AND SUPERVISORY OFFICERS**

1. The division of classes into two or more sections according to ability.
2. The selection of pupils for special classes, such as classes for the exceptionally bright or exceptionally dull pupils or for pupils having special defects in certain subjects.
3. The determination of the efficiency of the school as a whole by comparison of obtained scores with norms and with scores made by other schools or grades.
4. The determination of whether the proper emphasis is given to all subjects or whether some subjects are overstressed.
5. The comparison of different methods of instruction or comparison of new methods with the ones already in use.
6. The determination of the general achievement level of a grade, a school, or a system.
7. The measurement of the progress of a grade, a school, or a system, for a semester, a year, or any given period.
8. The determination of whether or not the grade, the school, or the system is achieving what can fairly be expected.

9. The determination of the misplacement of pupils in grades or sections.
10. The proper classification of new pupils entering the school system.

When to Give the Tests. Tests may be used periodically or they may be used somewhat constantly in the classroom. The type of testing followed depends somewhat upon the purposes the tests are to serve and the nature of the tests selected. If the tests used are the more common survey tests of general achievement, they are usually given early in the school term and then again a few days before the end of the school term. This procedure permits the teacher to determine the improvement which his pupils have made during this period. Tests which are used definitely for survey purposes and are given only once during the school year are frequently administered at or near the end of the year. This is probably one of the least important times for tests to be given, since almost the entire school year is gone and there is no opportunity for the teacher to attempt to do anything about the conditions revealed by the tests. *If only a single cross-section of the school is taken, this should undoubtedly come early enough in the school year to permit the teacher to profit from the findings.* The periodical use of educational tests to measure class or individual pupil improvement is by far the most prevalent practice.

A further refinement of the idea of using tests early in the school year is found in their use immediately following the completion of the instruction on a particular unit of subject matter. Unit achievement tests, each designed to measure a specific area of the course, are proving popular for this purpose with both teachers and pupils. By using these narrow-function tests immediately after the completion of the teaching of a specific subject-matter unit the teacher secures immediate information about the weaknesses of his class. The special inadequacies of his instruction are thus made clear, and he can proceed at once to set up a remedial program before the class has moved on to other subject matter. This suggests the continuous use of tests as the basis for remedial work. The information provided by the use of these numerous narrow-function tests is also valuable in organizing future instruction to prevent the appearance of these recognized weaknesses.

Cooperative Testing Programs. During the last decade, cooperative testing programs have developed in various cities and states for the purpose of providing a coordinated attack upon measurement and evaluation problems. These programs are very different in organization, sponsorship, and objectives,³ but they typically provide testing services of such a nature that the participation of most teachers is limited either to the administration of the tests and use of results or alone to the use of the results.

City Testing Bureaus. Bureaus of testing and measurement in a number of the larger cities maintain staffs of measurement and research specialists whose primary functions are to carry on planned testing programs and perhaps also to conduct related research studies. Frequently the cooperation of teachers is obtained in the administration of tests, and the results are made available to them for use with their pupils. Programs are frequently planned in cycles of several years, and tests in line with the total program may be given annually, twice a year, or at more frequent intervals.

State-Wide Testing Programs. Testing and related services are now available to the schools of approximately half of the states through some public educational agency in each state. Frequently the testing programs are based on cooperative construction, administration, and scoring of the tests and uniform methods of reporting results in comparable form. In other cases, available standardized tests are cooperatively administered and scored and the results are reported in as uniform a manner as possible. State-wide norms are frequently provided. New forms of tests are constructed or provided annually in some state-wide setups.

These programs and services vary widely among the different states, and include various patterns of achievement, intelligence, and personality tests. Some of the programs are conducted as scholarship contests, some are cooperatively sponsored by collegiate institutions which make use of scholastic aptitude test results of high school seniors, some are conducted primarily for purposes of supervision, and still

³ David Segel, *National and State Cooperative High School Testing Programs* U S Office of Education Bulletin, 1936, No 9 Government Printing Office, Washington, D C, 1936

others are administered purely as services to the schools. Schools in some states participate in the programs on a cost basis, and participation is most often optional for each school.

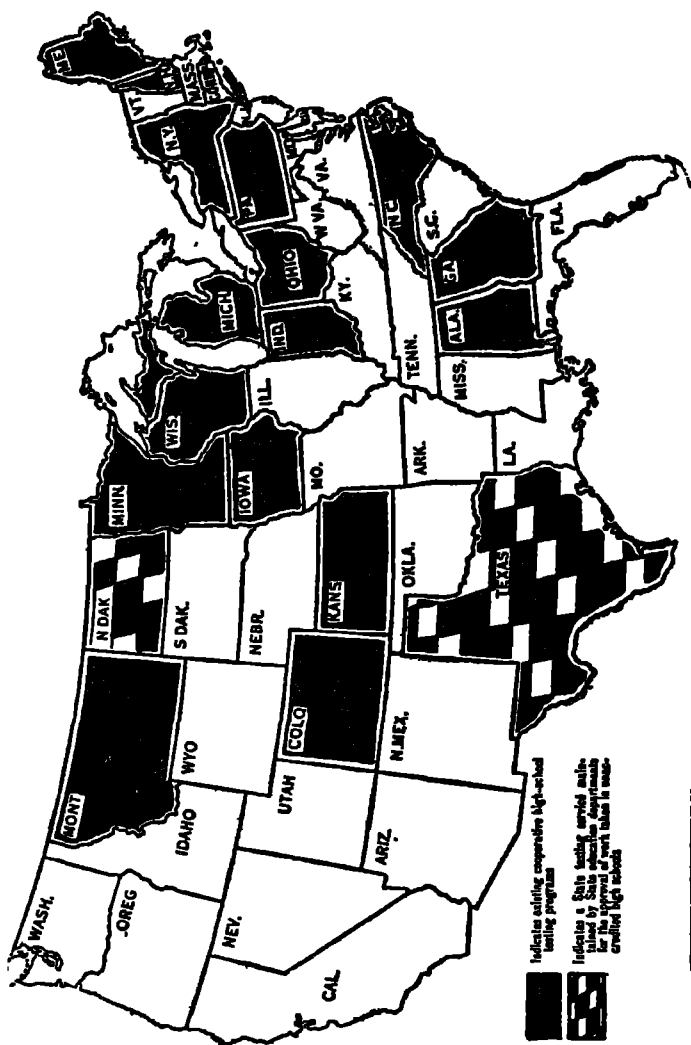
The accompanying map indicates the states which Segel, in the investigation mentioned above, found to be conducting such programs in 1933. State-wide testing services have since been inaugurated in at least several states and cooperative arrangements with nearby programs have permitted some states not themselves providing such programs to participate in such cooperative testing, so it seems certain that at least half of the states are now serviced directly or indirectly by such programs. Although the elementary school level is probably not so well provided for, it is known that some of the states also have cooperative testing services for the elementary grades.

Nation-Wide Testing Programs. Cooperative testing on a nation-wide basis is offered by various educational foundations, cooperative services, and commercial agencies for a wide variety of educational and mental tests. The services are sometimes provided primarily for a particular group of schools and in other cases are provided for any school wishing to obtain them. Reports are furnished to participating schools, and norms are often prepared on regional and nation-wide samplings of pupil test results.

III. SELECTING THE TESTS

Test selection depends upon the type of testing planned, as a basic consideration, for tests should be chosen which not only are within the proper subject field and at the appropriate level of advancement for the pupils but which also will serve the desired function. It should be pointed out that not all tests are appropriately named, however, and that too much dependence can easily be placed upon a test title. Accordingly, the student and teacher should learn to utilize critical standards in the selection of testing instruments.

Need for Care in Selecting Tests. As has been implied above, the mere fact that a test is standardized guarantees neither its validity for the type of use prescribed by its author and publisher nor its validity for the use to which a

FIGURE 7 STATES HAVING STATE-WIDE COOPERATIVE TESTING PROGRAMS, 1933⁴⁴ Ibid. p. 33.

teacher wishes to put it. A valid test of achievement should consist of items which are in harmony with facts and accepted usage in the subject in question. Yet, it has been found that nearly ten percent of the items in sixteen general science and biology tests contain errors⁵ and that in five tests of English usage from one-sixth to more than one-half of the usages scored as wrong in the different tests are acceptable in terms of the standards acceptable to the National Council of Teachers of English.⁶

The teacher or administrator selecting standardized tests has a right to expect that accurate information will be furnished him by the author and publisher concerning the validity, reliability, and other criteria of a good examination. A source of evidence about tests is the *Mental Measurements Yearbooks*, published in new editions at frequent intervals.⁷ These yearbooks contain carefully edited descriptions and critical reviews of tests by subject matter and test specialists with which to supplement information about a test furnished by the author and publisher.

Test Rating Scales. In the discussion of criteria for tests in a previous chapter, no attempt was made to evaluate in a definite manner any of the items which appear to affect test quality. Numerous rating devices which weigh the various items roughly in order of importance are available. The assignment of point values to the different features of the tests is, of course, purely a subjective procedure, as is to a great extent the actual use of such score cards by different individuals. It is obvious that two different individuals using score cards could not be expected to agree closely in the point scores assigned to a particular test. However, in spite of these limitations, such rating scales are of very real value to the inexperienced teacher or student because of the definite way in which attention is called to the quality features in a test.

⁵ Leon N. Diamond, "Testing the Test-Makers" *School Science and Mathematics*, 32 490-502, May 1932.

⁶ Karl W. Dykema, "On the Validity of Standardized Tests of English Usage." *School and Society*, 50 766-68; December 9, 1939

⁷ Oscar Krisen Buros (Editor), (1) *The 1938 Mental Measurements Yearbook*. Rutgers University Press, New Brunswick, N. J., 1938; and (2) *The Nineteen Forty Mental Measurements Yearbook*. The Mental Measurements Yearbook, Highland Park, N. J., 1941.

Because of the very satisfactory and detailed analysis it furnishes of the factors which should be considered as criteria for tests, the *Cole-von Borgersrode Scale for Rating Standardized Tests* is reproduced here. The student will find that time spent in the study and application of this scale in the actual evaluation of educational tests will be a most effective method of giving definite meaning to the previous discussion.

COLE-VON BORGERSDODE SCALE FOR RATING STANDARDIZED TESTS⁸

I. Preliminary Information

1. Exact name of test ?
2. Name and position of author ?
3. Name of publisher and nearest address ?
4. Cost ?
5. Date of copyright ?
6. Purpose of test ?

II. Validity (25)

A. Curricular (15)

1. Exact field or range of education functions which test measures ?
2. Ages and grades for which intended ?
3. Criteria with which material was correlated ?
4. Do questions parallel good teaching procedures ?
5. How wide is sampling of important topics ?
6. What is the social utility of questions ?
7. Is test claimed to be diagnostic ? (If so, proof, and see VI, 5, c, below)

B. Statistical (10)

1. Correlated against what outside criteria ?
2. Size of coefficient of correlation ?
3. Size and representativeness of sampling ?
4. Proof of validity of items ? (such as statements as to experimental tryout of items individually to determine that no large percentage is failed or passed by all pupils and that the items show a consistent increase of percentages of successes with successive age or grade levels).

III. Reliability (25)

A. Most important items

1. Correlated with what ?
2. Size and representativeness of sampling ?

⁸Robert D. Cole and Fred von Borgersrode, "A Scale for Rating Standardized Tests" *School of Education Record of the University of North Dakota*, 14:11-15, 1928

3. Reliability coefficient ?
4. The means of the distributions ?
5. The standard deviations of the distributions ?
6. If some other measure than the above three is given to prove reliability, what is it ?
7. Inter-correlations ?
- B. Less important but desirable
 1. Order of giving various forms of test ?
 2. Is test reliable enough statistically for individual measurement, or can it be used only for groups ?
 3. Evenness of scaling ? (see II, B, 4)
 4. Are pupils accustomed to this type of test ?
- IV. Ease of Administration (15)
 1. Manual of Directions (3)
 - a* How complete and simple is the manual ?
 - b* Does manual control test conditions well ?
 - c* Typographic make-up ?
 2. Simplicity of Administration (8)
 - a*. Amount of explanation needed for pupils by examiner ?
 - b*. Are directions to pupils clear, detailed, comprehensive ?
 - c*. Is arrangement of test convenient for pupils ?
 - d*. Are samples and "fore-exercises" given when needed ?
 3. Alternate forms (3)
 - a* Number ?
 - b* Evidence of reliability ?
 - c* Evidence of equivalency ?
 4. Time needed for giving (1)
- V. Ease of Scoring (10)
 1. Degree of objectivity — purely objective or some judgment on part of examiner ?
 2. Are adequate directions given — clear, equal to all emergencies ?
 3. Is scoring key adjusted to size of test ?
 4. Time needed to score one test ?
 5. Simplicity of procedure ?
 - a* Number of processes needed to get final score ?
- VI. Ease of Interpretation (20)
 1. Norms (6)
 - a* Kind — age, grade, percentile, etc ?
 - b* Derivation — size and representativeness of sampling ?
 - c* Tentative, arbitrary, or experimental ?
 - d*. For separate parts ?
 - e*. How expressed ?
 2. Is class record provided ?
 3. Are there provisions for graphing results ?
 4. Is interpretation of raw scores easy or hard ?

5. Application of results (10)
 - a. Are directions or suggestions given for application of results to benefit teaching or administration?
 - b. Are tests survey or diagnostic?
 - c. If diagnostic —
 - (1) Proof of diagnostic value?
 - (2) What principle or principles underlie construction?
 - (3) How many different skills, abilities, or aspects of the subject are analyzed or measured?
 - (4) Does the analysis of total subjects into unit abilities follow teaching practices or needs?
 - (5) Is the diagnosis individual or class-proof?
 - (6) Does the test demand tabulations of individual pupils' errors to secure diagnosis?
 - (7) Is a remedial program provided or suggested?

VII. Miscellaneous (5)

1. Typography and make-up?
 - a. Arrangement of printed matter?
 - b. Legibility of type?
 - c. Quality of paper?
 - d. Are test blanks free from distractions, norms, directions to examiner, etc.?
2. Is the time required for giving as small as is consistent with reliable measurement?
3. Is the cost in keeping with the amount, scope, and reliability of the results yielded?
4. Is good test service provided by the publisher?
5. Kind of new-type questions used?

The accompanying reproduction of the *Otis Score Card for Rating Standardized Tests* indicates the weights assigned to the various criteria of a good examination. The student may wish to note the major importance assigned to validity, reliability, and ease of administration, scoring, and interpretation, which parallel the criteria of a good examination emphasized in the treatment of that subject in Chapter IV above.

IV. ADMINISTERING THE TESTS

The general procedures suggested below are common to most tests now in use. They are not intended to take the place of the directions accompanying the various tests that may be used. The directions for giving and for scoring

OTIS SCORE CARD FOR RATING STANDARDIZED TESTS⁹

Item	Standard Number Points	Names of Tests			
1. Manual	7				
2 Validity	20				
3 Reliability	10				
4. Reputation	3				
5. Ease of Administration (20)					
<i>a.</i> Little special preparation	4				
<i>b.</i> Adequate detailed directions	6				
<i>c.</i> Time limits clearly stated	6				
<i>d.</i> Alternate forms available	4				
6. Ease of Scoring (15)					
<i>a.</i> Objectivity	8				
<i>b.</i> Convenient form of Key	4				
<i>c.</i> Time required	3				
7. Ease of Interpretation (20)					
<i>a.</i> Types of norms	10				
<i>b.</i> Directions for	3				
<i>c.</i> Class Record Sheet	2				
<i>d.</i> Remedial Program	5				
8 Typography and Makeup	5				
Total	100				

⁹ *Scale For Rating Tests* Test Service Bulletin, No 13. World Book Co., Yonkers-on-Hudson, N. Y., 1926.

supplied in the examiner's manuals which accompany the better tests should be rigorously followed in order to guarantee that the tests are given under standard conditions.

Preparation for Testing. Any teacher or principal who is reasonably skillful in discipline, and who will carefully follow the directions accompanying the tests, should be able to administer a modern educational test. Unless the test directions are extremely familiar, the examiner should study the manual carefully before attempting to give the test. If possible he should administer the test to some other person in order to gain further familiarity with the procedure. If this is not possible, the directions should be read aloud several times so that they may be followed easily as the test is given. Familiarity with the directions is essential if the standard conditions for the test are to be maintained, and valid comparison of results with the norms thus be made possible.

Pupils may be tested in ordinary classroom groups or in larger groups. If several grades are to be given the same test, time may be saved by moving all pupils into a larger room, care being taken that the seats and the desks are suitable.

Before the test folders are given out, the desks should be cleared and each pupil should be provided with a sharpened pencil, or, if the test is to be scored by machine, with a special electrographic pencil. A number of extra pencils should be available for emergencies during the examination. The room should be quiet throughout the test. No questions should be allowed during the test. A manner which is agreeable but which at the same time suggests authority should be cultivated. Pupils should be made to feel "at home" in taking the test. Pupils will look forward to taking tests without fear or nervousness if the tests are properly given and if no misconceptions with regard to the meaning and use of the results are allowed to arise.

Administration of the Tests. Throughout the examination, directions should be given in a forceful manner, and should be spoken slowly and with careful attention to emphasis. The voice should be just loud enough to carry to all parts of the room. The directions accompanying the

tests should be followed verbatim. As far as possible, disturbances within or without the room which might interfere with the administering of the tests should be prevented. To avoid interruptions, the teacher may prepare a card carrying these words: *Testing Going On. Please Do Not Disturb.* If this card is hung on the outside of the classroom door, interruptions will be less frequent.

The time limits as set in the directions for giving the tests should be strictly observed. Tests should be timed to the second, or the results may not be comparable to what others get when the exact time is taken for the test. In timing the test, a stop watch is very desirable. If an ordinary watch is used, one having a second hand, so that the minute and second hands can be synchronized, is preferable. The following illustrative procedure will serve quite well if a stop watch is not available:

	Hr.	Min.	Sec.
(a) Record time starting signal is given	11	18	20
(b) Add to this the time required for the test	—	15	00
(c) The sum is the time to signal a stop	11	33	20

The Teacher's Responsibility. In the earlier stages of the development of standard tests, it was believed that the most valuable results came from their use in a periodical survey by persons other than the classroom teacher. More recently it has come to be more generally accepted that as many of the tests as possible should be given by the classroom teacher. This seems to be especially true in the case of tests which furnish information of special importance in the improvement of classroom instruction. In addition to allowing the classroom teacher to become acquainted with the technique of testing, it gives him a first-hand opportunity to observe the reactions of his individual pupils in the various test situations. On this account, it is believed that, wherever test results are to be used definitely as a basis for the discovery of individual pupil difficulties, tests should as far as possible be administered by the classroom teacher himself. However, where the test results are used for a survey of achievement in the entire school or system, it is less important for the teacher to have an intimate contact with the testing

program. As a matter of fact, many school administrators prefer not to have the teachers give the tests when they are used for such survey purposes.

V. SCORING THE TESTS

The scoring methods and devices discussed below are those used more or less widely with various standardized tests. Other procedures which are not especially designed for specific standardized tests but which are more widely used for the informal objective examination are discussed in Chapter VIII.

Hand-Scored Tests. The scoring of most standard tests is made almost wholly objective by the use of hand-scoring keys. The answer keys and directions for scoring each specific test should be followed rigorously. Scores should be obtained in exactly the manner prescribed by the test authors, in order that they may be compared directly with the norms which have been derived for the tests. It is best that all calculations be performed twice, and that all transcribed records be checked against the pupils' test papers to make sure that no errors have been made.

Hand-scoring keys of several types are commonly used, among the most common being strip keys, cutout stencils, and transparent stencils. When answers are given in column form, strip keys which have correct answers spaced on narrow strips of cardboard to correspond in spacing with the items of the test may be placed alongside a pupil's work for rapid scoring. When answers are scattered over a page and whenever the answer itself is the only point requiring the attention of the scorer, stencils having correct answers adjacent to apertures cut so they will fall directly over the pupil's answers when the key is placed over the test also permit of rapid scoring. Transparent stencils are similar to the above type, but they do not usually permit the scorer to check the pupils' answers directly on his test paper.

The matter of responsibility for scoring hand-scored tests constantly arises as an administrative problem in the smaller schools. Teachers are likely to feel that the responsibility for scoring standard tests given for supervisory purposes

Metropolitan: ADV COMPL: Key-A				
TEST 7 HIST & CIV (Page 80)	TEST 8 GEOGRAPHY (Page 80) 20. (3)	TEST 8 GEOGRAPHY (Page 30) 64. (1)	TEST 9 SPELLING (Page 31) COLUMN 1 Grade 7	TEST 9 SPELLING (Page 31) COLUMN 1 Grade 8
74. (3)	21. (4)	65. (4)	1. thread	1. decide
75. (5)	22. (3)	66. (4)	2. pumps	2. prepared
76. (2)	23. (3)	67. (1)	3. minutes	3. missed
77. (1)	24. (2)	68. (1)	4. supply	4. earliest
78. (4)	25. (4)	69. (1)	5. happened	5. advised
79. (1)	26. (4)	70. (1)	6. excuse	6. patient
	27. (1)	71. (1)	7. fare	7. examination
80. (2)	28. (1)	72. (3)	8. suppose	8. crumb
81. (3)	29. (3)	73. (3)	9. ache	9. omitted
82. (6)	30. (2)	74. (1)	10. advertise	10. blizzard
83. (2)	31. (4)	75. (4)	11. decide	11. engineer
84. (6)	32. (3)	76. (3)	12. prepared	12. athletics
85. (1)	33. (3)	77. (2)	13. missed	13. carnival
	34. (1)	78. (4)	14. earliest	14. subscription
86. (2)	35. (1)	79. (4)	15. advised	15. sincerely
87. (1)	36. (2)	80. (1)	16. patient	16. session
88. (2)	37. (4)		17. examination	17. assistance
89. (3)	38. (1)		18. crumb	18. foreign
90. (4)	39. (3)		19. omitted	19. arrangement
91. (3)	40. (4)		20. blizzard	20. probably
	41. (4)		21. engineer	21. delicious
	42. (1)		22. athletics	22. quantities
			23. carnival	23. yacht
			24. subscription	24. delegates
			25. sincerely	25. opportunity

FIGURE 8 SAMPLE STRIP SCORING KEYS FOR METROPOLITAN ACHIEVEMENT TESTS ¹⁰

¹⁰ Richard D Allen, et al, *Metropolitan Achievement Tests*, Advanced Battery. Published by World Book Co, 1932.

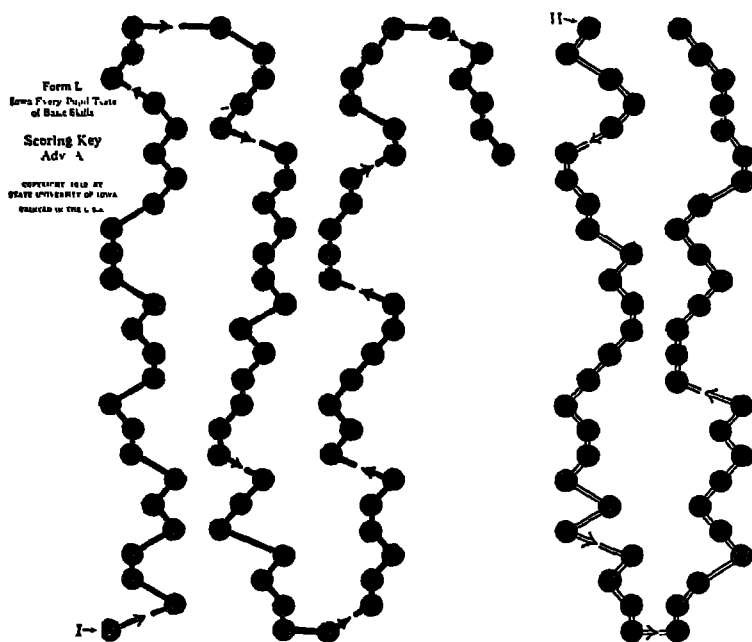


FIGURE 9 CUTOUT SCORING STENCIL FOR IOWA EVERY-PUPIL TEST OF BASIC SKILLS ¹¹

should not fall to them. A part of this difficulty arises through a failure on the part of the administrators to make perfectly clear to the members of the teaching staff their responsibility with regard to this type of work at the beginning of their terms of service. Most teachers, if given a suitable amount of time, do not seriously object to scoring test papers for their classes, particularly when they come to realize that this work is almost certain to reveal information which will be extremely significant to them in the improvement of their teaching practices. Teachers should consider it an opportunity, rather than an additional responsibility, to correct the test papers for their classes and to study the results of the tests. If a real interest in the outcome of the

¹¹ H. F. Spitzer, *Iowa Every-Pupil Test of Basic Skills, Advanced Battery, Test A Silent Reading Comprehension*. Published by Houghton Mifflin Co., 1940.

the test, is slit open for scoring. Correct answers appear in designated spaces on the back of the sheet for ready counting, while incorrect answers appear outside of the designated positions. Test folders are adapted for direct use with a number of standardized tests and are also available in a generalized form for use with informal objective examinations.

A machine scoring device which in effect makes self-scoring tests of instruments consisting of the common objective item forms is *Krexit*.¹³ Pupils respond to the items of a test by marking appropriate squares on special cards. The cards are inserted into a machine which is adjusted in advance for the test to be scored. When a lever is pulled, the correct answer positions are printed with red circles. Scores are obtained by counting the pupils' correct answers. The accompanying illustration shows a test card which has been marked by the machine.

Machine-Scoring Devices. The *International Test Scoring Machine*¹⁴ scores pupil answer sheets by means of an electrical current flowing through the lead deposited by the pupil's electrographic pencil on the answer sheet. Items of the alternate-response, multiple-choice, matching, and modified completion types can be scored by this method.¹⁵ Scores can be obtained by experienced machine operators at the rate of 700 or more per hour. Special answer sheets are provided and directly adapted for use with many of the newer standardized tests, while standard answer sheets in a variety of styles are available for the use of teachers or schools wishing to adapt their locally-constructed tests to machine scoring. The accompanying illustrations picture the test scoring machine and give examples of both types of answer sheets.

The *Multiplex Quick-Score Grader*¹⁶ is another machine for use in scoring objective tests. The pupil's test sheet is placed over a fibre board in which holes correspond to answer positions on the test sheet. The pupil punches holes

¹³ Manufactured by Krexit, Inc., Point Marion, Pennsylvania

¹⁴ Manufactured by International Business Machines Corporation, New York.

¹⁵ *Methods of Adapting Tests for Machine Scoring.* International Business Machines Corporation, New York

¹⁶ Manufactured by Multiplex Display Fixtures Co., St. Louis.

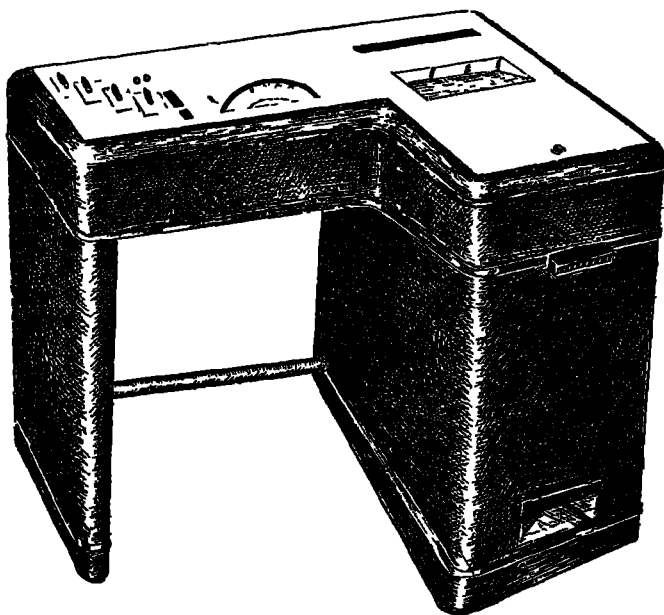
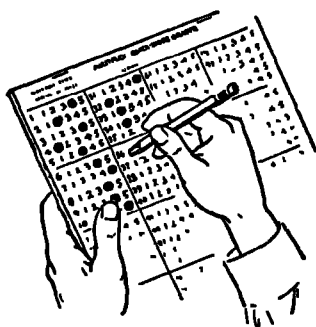


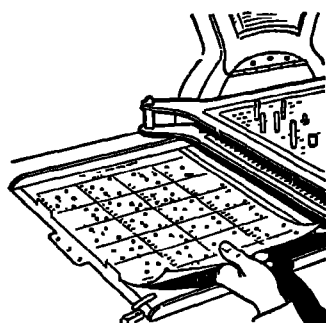
FIGURE 11. INTERNATIONAL TEST SCORING MACHINE

through the numbers indicating his answers. His test is then scored by "weighing" his responses. The test sheet is slid into a special rack which is suspended over the platform of a sensitive weighing scale. When the master plate, set for the particular test being scored, is lowered into position, small weights drop through the holes punched at the correct positions and contribute to the score which is read from a dial. The correct answers for items the pupil has missed can be indicated on the test sheet while the test is being scored. Items of the alternate-response, multiple-choice, and matching types can be used with this machine, and the manufacturer states that raw scores can be obtained at the rate of 300 to 600 per hour.¹⁷ Three styles of test sheets—for alternate-response and 3- and 5-response multiple-choice—are available. The accompanying illustrations show the methods by which the pupil indicates his answers on the test

¹⁷ "Grading Machine Automatically Scores Objective Tests." *School Executive*, 60.47, February 1941.



Taking Test



Scoring Test

FIGURE 13 ILLUSTRATIONS OF MULTIPLEX QUICK-SCORE GRADER

sheet and by which his test sheet is inserted into the grading device for scoring.

Use of Separate Answer Sheets. Prior to the development of machine scoring devices, separate answer sheets were used for hand-scoring by various persons with apparent satisfaction, and a few recent standardized tests have front covers which, torn from the booklet by the pupils taking the tests, serve as answer sheets for hand-scoring. Separate answer sheets for hand-scoring are used more frequently with teacher-made than with standardized tests, however.

An exhaustive study of the use of separate answer sheets reaches conclusions which in terms of test validities and reliabilities show that the separate answer sheets can justifiably be used.¹⁸ Dunlap reports that there is no evidence to show that separate answer sheets cannot be used with pupils in grades as low as the fourth. Some standardized tests provide separate answer sheet editions for pupils in the fourth and higher grades. However, for tests which are rather complex and which require a complicated answer sheet, it is preferable that separate answer sheets not be used much below the junior high school level.

¹⁸ Jack W. Dunlap, "Problems Arising from the Use of a Separate Answer Sheet" *Journal of Psychology*, 10 3-48, July 1940

VI. ANALYZING THE RESULTS OF TESTING

As a complete discussion of the statistical techniques used in analyzing scores resulting from the administration of tests is given in Chapter XXII, such questions will not be discussed here. However, it is pertinent to remark that the modern teacher is expected to understand and to be able to use such statistical techniques, so that he will be able to obtain maximum values in using the results from tests given to his pupils.

VII. INTERPRETING THE RESULTS OF TESTING

The results of testing are interpreted by the use of norms and also by the use of certain derived scores which are dependent upon norms. A section of the preceding chapter presented a discussion of the derivation, and to a certain extent the application, of norms for standardized achievement tests. Chapter XXIII presents a rather complete discussion of derived scores. Therefore neither norms nor derived scores are presented fully here. However, the educational quotient and other similar quotients are so definitely related to standardized achievement testing that they are treated briefly at this point.

The educational quotient (EQ) is a derived score which is made possible by the provision of age norms for many achievement tests. This quotient, which is similar to the better-known intelligence quotient, is based on the ratio between a child's educational age and his chronological age. His educational age is determined by the use of age norms for an achievement test battery, and is referred to frequently as an age equivalent. The educational quotient (EQ) is obtained by the use of the formula

$$EQ = 100 \frac{EA}{CA},$$

where EA designates the educational age and CA is the chronological age of the pupil. It is necessary that both educational and chronological age be stated in months in

applying this formula. For example, the EQ of a child of chronological age eight years who had an age score of nine years three months on a general achievement test would be

$$EQ = 100 \frac{9-3}{8-0} = 100 \frac{111}{96} = 116,$$

where his educational and chronological ages are both expressed in months.

Quotients of this type are not limited to general achievement, but are also possible in subject areas such as arithmetic, reading, language, science, and the social studies. To obtain a quotient comparable to the EQ in any of these areas, it is only necessary to know the age equivalent of a child's performance on a standardized test in that subject field and his chronological age. Thus, if a child's reading age is substituted in the above formula for his educational age, the result from the formula will be his reading quotient instead of his educational quotient. Such quotients are not as widely used as the IQ and the EQ, but serve useful purposes when a rather complete analysis of pupil achievement in a subject area is desired.

Quotients of this type indicate a child's educational achievement in relation to his life age. A quotient of more than 100 shows that he is achieving at a higher level than does the average child of his age, while a quotient of less than 100 is indicative of achievement at a level below that of the average child of his age. Deviations of considerable degree from 100 are found for pupils who are or probably should be accelerated in school and for pupils who are retarded in their achievement and perhaps their grade placement.

TOPICS FOR DISCUSSION

1. What should be the teacher's responsibility with regard to the use of standardized tests in the classroom?
2. From the standpoint of the classroom teacher, what are the major values of educational tests?
3. In what major ways are test results valuable in the classroom guidance of individual pupils?
4. Suggest a procedure by which properly designed tests may be used for individual pupil diagnosis. For class diagnosis.

5. What features of standardized tests make them particularly useful for pupil gradation?
6. Under what conditions are test results useful for comparative purposes?
7. Under what conditions do you believe the teacher should be responsible for the administration and scoring of standardized tests?
8. In your opinion, what types of tests (intelligence—individual or group—aptitude, general achievement, diagnostic or analytic, personality) should the teacher be encouraged to use most freely? Why?
9. What factors determine the time in the school semester or year when tests should be given?
10. Why is it desirable to have a clear-cut problem in mind in initiating a testing program?
11. Discuss the use of rating scales in the selection of standardized tests.
12. Discuss the steps you would follow in preparing for testing and in the administration of a standardized test.
13. What is the possible contribution of a state-wide or other type of cooperative testing program to the solution of local testing problems?

SELECTED REFERENCES

- Broom, M. E., *Educational Measurements in the Elementary School*, Chapter XIII. New York: McGraw-Hill Book Co., Inc., 1939.
- Broeckner, Leo J., and Melby, Ernest O., *Diagnostic and Remedial Teaching*. Boston: Houghton Mifflin Co., 1931.
- Buros, Oscar Krisen (Editor), *The 1938 Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938.
- Buros, Oscar Krisen (Editor), *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: The Mental Measurements Yearbook, 1941.
- The Cooperative Achievement Tests: A Handbook Describing Their Purpose, Content, and Interpretation*. New York: Cooperative Test Service of the American Council on Education, October 1936.
- Hawkes, Herbert E., Lundquist, E. F., and Mann, C. R. (Editors), *The Construction and Use of Achievement Tests*, Chapter IX. Boston: Houghton Mifflin Co., 1936.
- Hildreth, Gertrude H., *A Bibliography of Mental Tests and Rating Scales* (Second Edition). New York: The Psychological Corporation, 1939.
- Lang, Albert R., *Modern Methods in Written Examinations*, Chapters X, XII. Boston: Houghton Mifflin Co., 1930.
- Lee, J. Murray, *A Guide to Measurement in Secondary Schools*. New York: D. Appleton-Century Co., Inc., 1936.
- Lincoln, Edward A., and Workman, Linwood L., *Testing and the Use of Test Results*. New York: The Macmillan Co., 1935.

- Madsen, I. N., *Educational Measurement in the Elementary Grades*, Chapter X Yonkers-on-Hudson, N Y World Book Co., 1930
- Mead, A R, "Suggestions for the Training of Teachers in the Use of Educational Measurements." *Educational Administration and Supervision*, 12 23-43, January 1926
- Mort, Paul R, and Gates, Arthur I, *The Acceptable Uses of Achievement Tests*, Chapters I-IV. New York Bureau of Publications, Teachers College, Columbia University, 1932.
- Nelson, M J, *Tests and Measurements in Elementary Education*, Chapter XIII New York The Cordon Co., 1939
- Odell, C W, *Educational Measurements in High School*, Chapters XXI-XXIII New York The Century Co., 1930
- Orleans, Jacob S, *Measurement in Education*, Chapters 6-8. New York : Thomas Nelson and Sons, 1937
- Orleans, Jacob S, and Sealy, Glenn A, *Objective Tests*, Chapters IV-V, XII Yonkers-on-Hudson, N Y World Book Co., 1928.
- Ruch, G M, and Stoddard, George D, *Tests and Measurements in High School Instruction*, Chapters II-III. Yonkers-on-Hudson, N. Y World Book Co., 1927.
- Russell, Charles, *Standard Tests*, Part IV Boston Ginn and Co., 1930.
- Seder, Margaret, *Introduction to Testing and the Use of Test Results*. New York Educational Records Bureau, July 1940.
- Tiegs, Ernest W., *Tests and Measurements in the Improvement of Learning*. Boston Houghton Mifflin Co., 1939.
- Van Wagenen, M. J., *Educational Diagnosis and the Measurement of School Achievement* New York The Macmillan Co., 1926
- Woody, Clifford, and Sangren, Paul V, *Administration of the Testing Program*. Yonkers-on-Hudson, N. Y. World Book Co., 1933.

CHAPTER VII

USING ORAL AND ESSAY EXAMINATIONS IN THE CLASSROOM

The methods of using the oral and essay examination and the characteristics of these types of subjective tests mentioned below are the basis for the discussion of this chapter.

- a. Extent and importance of classroom testing.
- b. Limitations and advantages of the oral quiz.
- c. Place of the oral quiz in the schools.
- d. Limitations of the essay examination.
- e. Advantages of the essay examination.
- f. Improving the essay examination.

I. CLASSROOM TESTING

The problems involved in the construction, selection, and use of standardized tests have been discussed in the previous two chapters. This chapter and the following deal with the teacher-made or classroom test, as distinguished from the standardized test. From the point of view of many teachers, the classroom test constitutes the major problem of measurement.

Extent of Classroom Testing. Every teacher is faced with problems of measurement and evaluation in the classroom. Not all such problems are best solved by tests, for evaluation techniques of relatively subjective types have their place in the school and classroom. However, each teacher spends hours and days of time each year in preparing and scoring tests and in analyzing and interpreting the results. It has been estimated¹ that the approximately 700,000 teachers from the elementary school to the college level spend 42,000,000 hours of time annually, based on three hours per test and 20 tests per teacher, in testing. According to this general estimate, teachers average more than a week of school time annually in work with tests.

¹ William A. McCall, *Measurement*, p. 29. The Macmillan Co., New York, 1939.

Need for Improvement in Classroom Testing. The study of standardized tests thus far in this volume must make it apparent that by their very structure and use standardized tests do not meet all classroom needs for evaluation and measurement. In the first place, such tests do not equally well serve in all schools because of differences in emphasis and points of view resulting from the varying characteristics and educational needs of different communities. Also, classroom testing is sometimes important in an area so narrow or so specialized that no available standardized test fills the need. Again, teachers sometimes feel that standardized tests overstress factual knowledges and neglect what they believe to be important—the ability to organize and apply facts. For these reasons, written examinations prepared by local teachers, or at least within the local school system, will undoubtedly always be needed to meet the demands for complete and valid measurement of educational achievement.

Even a superficial observation of typical examination procedures of teachers makes it apparent, however, that the basic aims of examinations are not achieved in many instances, for the tests constructed and used by teachers frequently fail to accomplish what is expected of them. It is indeed unfortunate that teachers, sometimes not realizing that their tests fail to accomplish the desired purposes, unduly penalize pupils for lack of success on the tests. It is necessary, therefore, that the weaknesses of classroom tests be recognized and that steps be taken to bring teacher-made or classroom tests to as high a level of efficiency as possible.

II. THE ORAL EXAMINATION

Important as the oral quiz may be for instructional purposes, little need here be said concerning its use in the classroom for measurement purposes. As was pointed out in Chapter III, Horace Mann sounded the death knell for such a use of the oral examination nearly a century ago.²

² Otis W. Caldwell and Stuart A. Courtis, *Then and Now in Education*, 1845-1923, p. 37. World Book Co., Yonkers-on-Hudson, N Y, 1923

Limitations of the Oral Examination. To summarize Horace Mann's statements or implications, the oral examination : (1) is not equally fair and just to all pupils, (2) does not test extensively or efficiently, (3) permits of interference and favoritism, intentional or otherwise, by the teacher, (4) is unjustifiably time consuming, (5) leaves no permanent objective record of pupil performance, and (6) does not permit an evaluation of the difficulty of questions. While these indictments by Horace Mann accomplished very little in the sense of effecting any immediate widespread changes in examination practices, the weaknesses of the oral examination for measurement purposes have probably not since been stated more effectively.

Advantages of the Oral Examination. The oral examination or quiz does have some uses, however, in the evaluation and measurement of pupil performance, even though its values are seldom great when it is used in the classroom situation. Oral questioning can be effectively used in the Socratic manner to lead an individual pupil to certain logical conclusions through his own reasoning abilities; in this sense it is a teaching rather than a testing device. Oral questioning can be used with an individual pupil in probing his reasons for having responded as he did to certain questions on written examinations or on certain mathematical or scientific problems, in the attempt to determine the causes of error; in this sense it is a diagnostic testing tool. Oral questioning can be used in determining how well an individual pupil has integrated his knowledge, can apply it to various situations, and can see its implications; in this sense it is an evaluation tool, although admittedly a highly subjective one.

Oral questioning is widely used in individual intelligence testing, under rather rigidly standardized conditions and by a highly skilled examiner. Used in this manner it becomes a reliable technique for the measurement of mental ability. Oral questioning is the technique widely used in the interview; in this sense it serves the purpose of obtaining information quickly and accurately. It is thus apparent that oral questioning does have very real values for certain specific purposes and in certain specific situations.

In considering the above legitimate uses of oral question-

ing, it should be clearly noted that the conditions under which this method is properly used and the purposes it is appropriately expected to serve are very different from those operating when it is used with a group of pupils in the classroom to determine educational achievement. In general, the oral examination has no proper place in the classroom for measuring achievement, especially as a basis for determining pupil marks in a course.

III. THE ESSAY EXAMINATION

The traditional or essay examination has occupied, and today continues to occupy, an important place among the testing techniques used by the classroom teacher, although during the past few decades it has lost the dominant position it occupied at the turn of the century. Skepticism concerning the traditional examination arose more than a decade prior to 1900.³ Edgeworth published in England during 1890 what was perhaps the first critical study of the essay test.⁴ It remained, however, for Starch and Elliott to bring the issue sharply to the front in America in 1912 by a report of marks assigned to an English examination paper by various teachers⁵ and to follow it shortly by similar reports on two other subject fields. Although it is probable that educators for various reasons somewhat misinterpreted the findings of these and many subsequent studies of the traditional examination, the fact remains that the studies desirably called attention to a major weakness of this testing technique.

Limitations of the Essay Examination. Two major limitations and several related minor limitations characterize the essay examination. The two major limitations of the essay examination, (1) the factor of limited sampling, and (2) subjectivity of scoring, are discussed in some detail in the following paragraphs, and the minor limitations are discussed briefly.

³ I. L. Kandel, *Examinations and Their Substitutes in the United States* Bulletin No. 28, Carnegie Foundation for the Advancement of Teaching, no. 27-35. The Foundation, New York, 1936.

⁴ F. Y. Edgeworth, "The Element of Chance in Competitive Examinations" *Journal of the Royal Society*, 53 460-75, 644ff., 1890.

⁵ Daniel Starch and Edward C. Elliott, "Reliability of Grading High School Work in English." *School Review*, 20 442-57, September 1912.

Limited Sampling. The first major limitation of the essay examination is its limited sampling. A test which consists of five or ten questions cannot hope to sample widely over any sizable field of subject matter, but can measure only over a few of the important areas in which pupil abilities should be tested.

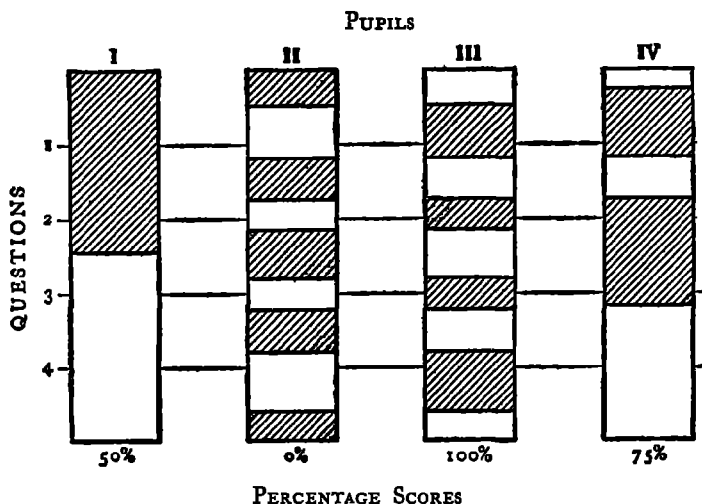


FIGURE 14 THE INFLUENCE OF LIMITED SAMPLING ON TEST SCORES

Figure 14 shows in graphic form an hypothetical testing situation which brings out the undesirable results of limited sampling. Pupils I, II, III, and IV each knows exactly half of the material over which the test is to be given. However, the particular facts mastered by each pupil are not the same throughout. For example, Pupil I, who was perhaps regular in his attendance during the first half of the course, has a mastery of the earlier units of the subject-matter. This is indicated by the shaded portion of the column. The second pupil, through irregular attendance, spasmodic preparations, or other unknown causes, mastered a few of the facts, missed another section, then learned a few more, etc. Pupil III was just as irregular in his attendance, but for some reason learned exactly those items missed by Pupil II. Pupil IV shows merely another variation of the situation.

It might be carried on almost indefinitely, but four cases are adequate to illustrate the entire range of variation due to sampling.

Now if a brief essay-type examination consisting of four questions from the various areas of the course is given, it will be noted that distinctly different types of response are secured from these four pupils. Pupil I, knowing the facts in the first part of the work, responds to the first two questions and makes a score of 50 percent. The second pupil, by sheer bad luck in the selection of the facts he learned, misses each of the four questions, and receives a zero score. Pupil III, through good fortune (or judgment), happens to

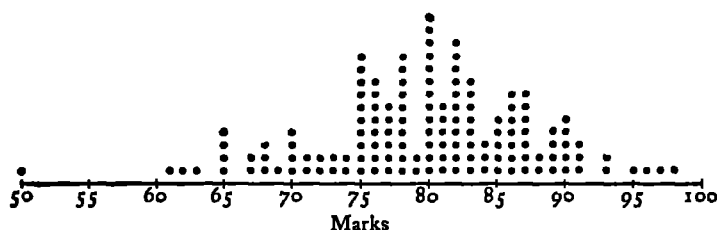


FIGURE 15 MARKS ASSIGNED TO AN ENGLISH EXAMINATION BY 142 TEACHERS⁶

have mastered the items in the exact areas sampled by the test, and thereby makes a perfect score on the test. Pupil IV, as another variation due to chance, scores 75 percent on the examination. Thus there is a variation of from 0 to 100 percent on the examination taken by four pupils each of whom actually has a mastery of exactly 50 percent of the facts. This type of error in measurement of achievement, which unfortunately is not uncommon, is due to the factor of limited sampling.

Subjectivity of Scoring. A second outstanding characteristic of the essay examination is subjectivity of scoring. Starch and Elliott, who had 142 teachers score identical copies of an English examination paper, found that the scores based on 100 percent for perfection ranged from a low of 50 to a high of 98.⁷ In another study, they found

⁶ Adapted from Starch and Elliott, *op. cit.*

⁷ *Ibid.*

that 115 teachers rated a geometry paper from a low of 28 to a high of 92.⁸

Starch and Elliott's studies were followed during the next ten or so years by many similar investigations by other research workers. Ruch, for example, had 91 teachers of geography score the essay examination papers judged to be the best, average, and poorest papers from a class on the basis of 20 for an entirely satisfactory answer and 0 for an answer practically without discernible merit. The range of scores on the best paper was from 3 to 20, on the poorest paper from 0 to 2, and on the average paper from 2 to 20, with average scores being 16.1, 0.1, and 10.9 respectively for the best, poorest, and average papers.⁹

Eells¹⁰ had 61 teachers score an examination consisting of four essay questions in geography and history, and eleven weeks later had them score the same answers again. Reliability coefficients, obtained by correlating the first and second scores assigned by the same teachers, ranged from 0.25 to 0.51 for the four essay questions. This and other evidence showing wide differences in the two sets of scores assigned by the same persons, led him to conclude that the same individuals vary from time to time in their judgments about as widely as different individuals vary.

Although some dissenting voices were raised and evidence obtained by some investigators using similar techniques showed relatively small differences among the scores assigned by various teachers,¹¹ the conclusions usually reached by studies of this type were that the scoring of an essay examination is a highly subjective process and that the resulting scores are correspondingly inaccurate.

The effect of a lack of objectivity in the unit of measurement may be demonstrated to anyone who will try to measure the length of a table top by using a rubber band as the measuring instrument. The length of the table in rubber band

⁸ Daniel Starch and Edward C. Elliott, "Reliability of Grading High School Work in Mathematics" *School Review*, 21 254-59, April 1913

⁹ G. M. Ruch, *The Objective or New-Type Examination*, pp. 78-81. Scott, Foresman and Co., Chicago, 1929

¹⁰ Walter C. Eells, "Reliability of Repeated Grading of Essay Type Questions" *Journal of Educational Psychology*, 21 48-52, January 1930

¹¹ F. E. Bolton, "Do Teachers' Marks Vary as Much as is Supposed?" *Education*, 48 23-39, September 1927

units depends on how much tension is placed on the rubber band. Obviously no accurate measurement could result.

The subjectivity of scoring shown by the above studies of the essay examination is the result more of standards of expectancy which varied widely among the teachers concerned than of any other cause. Such standards of expectancy vary from teacher to teacher, grade to grade, and school to school. Unfortunately, from the point of view of improving the accuracy of scoring the essay test, this limitation appears to be largely innate in the type of examination itself. The establishment of uniform standards of achievement in the teacher is probably a human impossibility. The remedy lies not in the attempt to produce it but in giving the teacher a tangible unit of measurement.

TABLE IX
SHIFTING STANDARDS OF EXPECTANCY

Quality of Product	Grade				
	4	5	6	7	8
82	A	A	A	B	C
68	A	A	B	C	D
50	A	B	C	D	F
32	B	C	D	F	F
18	C	D	F	F	F

Shifting standards of expectancy may be illustrated by the data of Table IX. Here are shown the shifts in standards which enter into teachers' estimates of school products. If it is assumed that a given school product, such as a handwriting or drawing specimen, has a rating-scale value of 50, it appears from the table that the specimen would receive a mark of *A* at the fourth-grade level. It would represent distinctly superior work for that grade. To the eighth-grade teacher, however, the specimen would appear to be very inferior as an **eighth-grade** product and the very poor mark of *F* would be assigned to it.

Minor Limitations. Another factor which affects the teacher in his marking of examination papers, but which did not enter into the studies reported above, is that he typically knows the pupils whose papers he is marking and also ordi-

narly knows whose paper he is scoring at a given time. He is certain to be influenced by that knowledge. He is probably prone to give pupils who have previously done good work or at least made favorable impressions on him the advantage of the "halo effect." This term describes the tendency to give high marks to such pupils in some instances where they are not deserved, by explaining to himself, perhaps unconsciously, that he *knows* they know the correct answers even though their responses to the questions are not highly satisfactory to him. Similarly, the pupils he has catalogued as of low ability are sometimes penalized by his tendency to consider their good answers merely "shots in the dark" or as implying ideas which the pupils did not actually understand.

Still another type of factor affecting the objectivity of scoring of an essay test is found in the influence upon the reader of handwriting and general neatness of the paper; spelling, punctuation, and grammar, organization of the paper; and even its length. It is certainly true that such characteristics as good handwriting, English usage, and organization predispose the reader toward high marks, and it is self-evident that the slow writer is penalized if a premium is attached to length of responses apart from their quality in examinations where the time is rigidly restricted. Some teachers penalize a pupil for deficiencies of these types, but other teachers do not. Moreover, the same teacher may penalize a pupil for such deficiencies one day and not do so on another occasion, depending upon his mood at the moment, and may penalize some pupils and not others.

Other influences which in considerable degree enter into the marking of tests are evidences of pupil effort, improvement, attitude toward the teacher and the course, conformance, and a multitude of other indications of what the teacher might consider desirable behavior on the part of the pupil. Some teachers believe in assigning relatively higher marks to pupils who try but do poorly than to pupils who appear not to try but do well. Others assign good marks to pupils who conform to sometimes inconsequential and irrelevant demands and penalize pupils who do not conform. The two types of scoring errors accounting for the subjec-

tivity of the essay test are known as constant errors and variable errors. Constant errors are those which result from a tendency to mark high or to mark low, i.e., to be an "easy" marker or a "hard" marker. Variable errors result from the tendency of all persons to vary in their judgments from time to time, according to their states of mind, the states of their digestions, and many other factors.

Pupils who do not know the answers to essay questions are prone to respond in terms calculated to cover up their lack of information if not actually to mislead the teacher. Such responses, which tend to vary in plausibility directly in relation to the intelligence of the bluffer, may take the form of discussion concerning content closely related to that covered by the question, of very incomplete answers which by repetitious statements and copious illustration may give a sense of completeness, and various other devices. Whether bluffing is or is not desirable is not the issue. Certainly bluffing is resorted to in great or small degree by all persons on some, if not many, occasions. To the extent, however, that bluffing is actually successful on essay tests, the examination results are less accurate measures of pupil achievement.

Advantages of the Essay Examination. Only the major and rather commonly accepted advantages of the traditional or essay examination will be discussed here. It should be remembered here particularly that it is the total effect of the examination which is important rather than the specific aspects considered singly. An advantage, then, may not be an advantage when there is balanced against it one or more dependent disadvantages.

Ease of Construction and Administration. Essay tests are easy to prepare and administer, in general, for teachers and pupils commonly know the nature of essay test questions and the traditional methods of answering them. Teachers typically prepare essay questions in a minimum of time, sometimes immediately before an examination is to be given. Some even prepare the last part of the test while the pupils are writing on the first question. Little or no time need be taken for telling pupils how to take the essay examination. However, essay tests prepared and administered with a minimum of effort are likely to have such resulting disadvantages

that the saving of time and labor may well be at the expense of testing efficiency.

Adaptability to Subject Matter. It is possible to use the essay examination for practically all subjects of the school curriculum, for the question and answer method is widely adaptable. Some types of outcomes, such as arithmetic skills, handwriting skills, reading ability, and others, cannot be tested directly by this device, but the factual background for them frequently can be so tested. As a matter of fact, the essay test procedure is often used in scoring arithmetic examples by the use of arbitrary decisions in scoring for correctness of the result or correctness of the method, in giving partial credit for answers not entirely correct, and in various other ways

Measurement of Higher Mental Abilities. Advocates of the older type of examination insist that the discussion-type questions have values not possessed by the informal objective test in that they call for comparison, for interpretation of facts, for criticism, for defense of opinion, and for other types of higher mental activity. Essay questions allow for some range of choice, which makes possible the meeting of differences in courses and readings pursued. The purpose of the written test is primarily to ascertain, not if the student knows this or that in particular, or has done routine work well, but if he knows accurately a considerable amount and understands a considerable variety of matters in terms of their interacting relationships. What is sought is a measure of accurate knowledge of fact, understanding of difficult ideas, and reflection based on the ability to interpret and to criticize and decide. In short, the questions are devised to test the pupil's ability to make use of knowledge. This is particularly true for advanced students, for whom the testing of such types of higher abilities is more important than the testing of the broad factual knowledges which almost certainly have been acquired to a high degree.

Advantages Sometimes Claimed for the Essay Test. Various writers have claimed advantages for the traditional examination which were seriously considered some years ago but which are generally discredited today. There are also questions on which the evidence is so inconclusive or on

which the decision depends so much more on the philosophy of the teacher than upon definite research findings that positive advantages cannot be claimed with certainty.

Training in the Use of Written English. It has been contended by various persons in the past that training in the use of English is a logical function of the examination and that the essay test furnishes such training. However, neither contention is defensible. Courses in English provide training in the use of English, as do, indirectly and as by-products, many other types of school experiences. The examination, which has definite uses and purposes in the measurement area to occupy its attention, should not be expected to furnish training in the use of English, although, of course, written language is required in the essay test. Furthermore, the conditions under which the essay examination is typically given—pupils writing at high speed and without time to organize their thoughts carefully—are not conducive to the best use of English. Certainly examinations in such courses as language, composition, spelling, and perhaps reading and literature might be devised to furnish training in the written use of English, but there seems to be no justification for shortening the time given to the measurement of direct course outcomes in the sciences, social studies, arithmetic, etc., in order to furnish the pupils this type of training.

Motivation of Desirable Methods of Reviewing. It has been shown that some student groups prepare for essay tests more often by reviewing broadly the important aspects of course content but that they more frequently review for the objective test by memorization of facts, of exact wordings of the textbook, etc.¹² No one should attempt to deny the desirability of the first rather than the second type of review. However, the studies summarized by Meyer were few in number, they dealt with only a few groups of pupils, and for the most part they were based on questionnaire responses, i.e., how pupils say they prepare for examinations. Probably the *type* of examination is less important to the pupil in determining how he should review than the *nature* of the test. An essay examination may or may not stress detailed

¹² George Meyer, "The Essay Type of Examination" *American School Board Journal*, 89 17-18, December 1934.

facts. An objective test may or may not stress detailed facts. Teachers differ markedly in the emphases they assign to factual learning and to applications of facts in the tests they give.

Conclusions Concerning the Essay Examination. For many years the essay-type test has been subjected to intense criticism. In spite of these attacks, however, it is still in use in numerous classrooms and doubtless performs a worthwhile function there. While it is true that when the essay test is subjected to a critical appraisal under research conditions many of the claims which have been advanced for it do not stand up any too well, it is also true that it performs certain functions in the classroom and for the pupils which the other more objective forms of tests fail to accomplish. Without doubt the essay-type test is firmly fixed in educational practice. It is a type of examination with which all teachers are familiar, and with all of its faults it undoubtedly possesses sufficient merit to warrant some attention to its improvement.

A productive revival of interest in the essay test has occurred during the last decade or so. It is now recognized that only a portion of the variability of marks assigned to an examination by *different* teachers, as in the Starch and Ellhott and other studies, can be attributed to the unreliability of the essay examination. Whereas the different teachers used in such studies had very different educational aims and standards of excellence, the teacher who scores an entire set of papers attempts to apply the same set of standards to all papers, and has the benefit of experience with previous papers as a basis for doing so. Furthermore, the teachers in those studies used no scoring rules save those which they developed individually, but the teacher who scores a set of papers usually applies more or less tangible and consistent scoring procedures.

A final summation of the limitations and advantages of the essay examination cannot be conclusive. Certainly the limitations of the test as it has been, and perhaps even today is, most widely used greatly outweigh its advantages. However, the possibilities of the essay examination have only recently come under careful scrutiny, so it may be that when the essay test is used with optimum efficiency and for

carefully selected purposes its advantages will outweigh its limitations.

IV. IMPROVING THE ESSAY EXAMINATION

Many suggestions for improving the essay examination have been made during the past few years. Most of the suggestions have to do with : (1) the selection of test content and the framing of questions, and (2) the scoring of the test results. The discussion below presents a few of the approaches to the improvement of the essay test by these two methods, but does not attempt to consider how the test may be improved in the specific subjects of the school curriculum.¹³

Types of Essay Questions. Monroe and Carter classified essay-type questions with respect to the types of mental activity each is designed to elicit in the pupil, and presented both descriptive statements concerning, and examples of, the twenty varieties they distinguished.¹⁴ The descriptive statements and illustrative questions below are from Odell's adaptation and supplementation¹⁵ of the questions from Monroe and Carter's list.

1. Selective recall—basis given. (Name the presidents of the United States who had been in military life before they were elected)
2. Evaluating recall—basis given. (Name the three statesmen who have had the greatest influence on economic legislation in the United States.)
3. Comparison of two things—on a single designated basis (Compare Eliot and Thackeray as to ability in character delineation)
4. Comparison of two things—in general (Contrast the life of Silas Marner in Raveloe with his life in Lantern Yard)
5. Decision—for or against. (In which in your opinion can you do better, oral or written examinations ? Why ?)
6. Causes or effects. (Why has the Senate become a much more powerful body than the House of Representatives ?)
7. Explanation of the use or exact meaning of some phrase or statement in a passage. (Explain the meaning of the expression "Sinai's climb" in the line "We Sinai's climb and know it not")

¹³ The bibliography at the end of this chapter lists some selected references on the improvement of essay testing in specific subjects

¹⁴ Walter S. Monroe and R. E. Carter, *The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students* University of Illinois Bulletin, Vol XX, No 34, University of Illinois, Urbana, 1923.

¹⁵ C. W. Odell, *Traditional Examinations and New-Type Tests*, pp 207-10. The Century Co., New York, 1928.

8. Summary of some unit of the text or of some article read. (Summarize in about one hundred words the advantages of the hot-air furnace.)
9. Analysis (Mention several qualities of leadership)
10. Statement of relationships (Tell the relation of exercise to good health)
11. Illustrations or examples (your own) of principles in science, construction in language, etc. (Give an original sentence in Latin illustrating the use of the infinitive in indirect discourse)
12. Classification—usually the converse of No. 11. (To what group of plants do the mosses and liverworts belong?)
13. Application of rules or principles in new situations (In what countries other than Brazil would you expect to find rubber plantations?)
14. Discussion (Discuss the Monroe Doctrine)
15. Statement of aim—author's purpose in his selection or organization of material. (What was the purpose of the author in having Athelstane return to life after he was apparently dead?)
16. Criticism—as to the adequacy, correctness, or relevancy of a printed statement, or a classmate's answer to a question on the lesson (Criticize "Macbeth was wholly indifferent to the superstitions of his time")
17. Outline. (Outline in not more than one page the chief events of the French and Indian Wars)
18. Reorganization of facts (Select the incidents which characterize Portia in *The Merchant of Venice*)
19. Formulation of new questions—problems and questions raised (If you were asked to state how much you could trust the viewpoint of a particular historian about whom you know little or nothing, what questions would you want to have answered concerning him?)
20. New methods of procedure. (How might the plot of *Julius Caesar* be changed to make it a comedy rather than a tragedy?)

Questions of the essay type are classified by Sims¹⁶ into three types. (1) simple-recall, (2) short-answer, and (3) discussion. The simple-recall questions, demanding a short response which can be accurately scored, require a number, a date, a place, an event, etc., in answer to *how many*, *when*, *where*, *what*, etc., questions. The short-answer questions, demanding statement, phrase, or sentence responses which can be rated quite objectively, require answers to *define*, *identify*, *list*, *find*, *state*, etc. The discussion questions, requiring responses of such complexity that objectivity of scor-

¹⁶ Verner Martin Sims, "Essay Examination Questions Classified on the Basis of Objectivity" *School and Society*, 35 100-2, January 16, 1932

ing is difficult, request answers to *discuss, explain, describe, compare, outline*, etc. As Sims points out, some essay questions are sufficiently definite that responses can be evaluated objectively but others are so general that responses can be rated with reasonable accuracy only by the use of definite scoring rules or some similar method.

Increasing the Objectivity of Scoring the Essay Test. Approximately a quarter of a century ago, Kelly conducted an investigation into the causes of variation in teachers' marks on examination papers.¹⁷ He found that the use of a rather definite set of rules resulted in greatly reduced variations in scores when the papers were rescored. More recently, Stalnaker obtained reliability coefficients ranging from .84 to .99 for the scores assigned to essay examinations in a variety of high school subjects by experienced teachers when scoring rules were used.¹⁸ These reliability coefficients show a highly satisfactory degree of scoring accuracy, especially when it is considered that only the lowest coefficient was under .90. Other studies of the results obtained when the essay test was scored under closely controlled conditions substantiate the conclusion that the traditional examination can be reliably scored if proper precautions are taken.

Sims proposed a rating method of scoring essay examinations.¹⁹ He suggested that the readers work out for themselves acceptable answers to the questions and then use the following procedures :

- a. Quickly read through the papers and on the basis of your opinion of their worth sort them into five groups as follows (a) very superior papers, (b) superior papers, (c) average papers, (d) inferior papers, (e) very inferior papers. (Remember that in a normal group you would expect to have approximately 10 percent of *very superior* and 10 percent of *very inferior* papers, 20 percent of *superior* and 20 percent of *inferior* papers, and 40 percent of *average* papers. Do not, however, try to conform rigidly to this rule. Your group may not be a normal one.)

¹⁷ Fred J. Kelly, *Teachers' Marks* Contributions to Education, No. 66 Teachers College, Columbia University, New York, 1914.

¹⁸ John M. Stalnaker, "Essay Examinations Reliably Read" *School and Society*, 46 671-72, November 20, 1937.

¹⁹ Verner Martin Sims, "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating" *Journal of Educational Research*, 24 216-23, October 1931.

- b. Re-read the papers in each group and shift any that you feel have been misplaced.
- c. Make no attempt to give numerical grades or to evaluate each question. Place each paper on the basis of your general impression of the total.
- d. Assign letter grades to each group; beginning with A for the very superior group, B for the superior group, etc.

Stalnaker believes that the use of optional questions reduces the reliability of marking the essay examination and recommends that all pupils be asked to "run the same race" by answering the same questions.²⁰

Wrightstone recommends that essay tests be designed to measure only one objective of instruction at a time, such as an interpretation of facts, that all scorers agree on a definition of the objectives and on certain standards of values, that an ideal answer be formulated and each part assigned a certain number of points, and that an eleven-point scale from 0 to 10 be used for each test unit.²¹

The following suggestions, by largely eliminating the personal judgment or bias of the scorer, have been found valuable for use in scoring essay-type exercises :

1. Examinations should be scored by the one who makes out the questions. He should know exactly what responses are desired, and should write out his answers to the questions in advance.
2. Each pupil taking the test should write his name on the back of the test paper and the scorer should disregard the name until the test is scored. This eliminates the subjective factor of being influenced or biased in judgment because of former contacts with the pupil, insofar as the teacher does not become aware through handwriting, manner of expression, etc., of the writer's identity.
3. The scorer should not mark off for misspelled words or poor sentence structure, paragraphing, handwriting, etc. Similarly, he should not increase the score for excellence in these things. However, such factors may be indicated or checked on the examination. The reason for this lies in the fact that the function of the examination is to measure the pupil's abilities in a course and not his ability to write or to spell. If it is desirable to test his ability to write, spell, or use correct written English, suitable tests can be obtained for these purposes.

²⁰ John M. Stalnaker, "A Study of Optional Questions on Examinations" *School and Society*, 44 829-32, December 19, 1936

²¹ J. Wayne Wrightstone, "Are Essay Examinations Obsolete?" *Social Education*, 1.410-15, September 1937.

4. Each separate item should be scored in all of the papers consecutively. This is preferable to the correction of each entire test as a unit, for it permits the scorer to concentrate on the answer to a single test exercise and better to judge the merits of the several pupil responses to the same question.
5. Each question should be rated on a scale of ten, twenty, or a given number of scoring points. The total score should be obtained for each pupil by adding the scores on the different questions only after all of the scoring has been done.

Whatever rules are followed, they will necessarily be arbitrary and not always wholly defensible. The significant point in the use of rules is that they provide for a reasonable uniformity in the handling of the papers of all the pupils and also furnish a guide for the control of the irrelevant factors that may affect the objectivity of the scoring.

Steps for Improving the Essay Examination. Three conditions appear to be necessary in bringing about improvement in the teacher-made examination of the essay type. These conditions are:

- (1) *The exact purpose of the examination must be understood by both the teacher and the pupil.* The emphasis of the essay examination should be definitely on thought, reasoning, and other types of mental activity as applied to the materials of the course. The main concern is with topics which involve interest-centers or relationships and problematical issues. Questions involving judgments, synthesis, and generalizations are admittedly difficult to evaluate, but they reveal aspects of pupil mastery and mind-quality probably not obtained otherwise.
- (2) *The content of the examination should be governed by its purpose.* In general, a test should parallel the objectives and pupil outcomes of the course. This means that there should be a proper balance of test content not only with respect to the subject matter but also with respect to the types of abilities to use and apply informations which are desired pupil outcomes. Essay-type questions have been generally open to the criticism that they are hastily and carelessly prepared. The advocates of the improved essay examination are

quite positive in their insistence that the preparation and selection of suitable essay-type questions should consume at least as much time as is required to score the answers. If this is done, the value and the accuracy of the scores obtained are almost certain to be increased.

- (3) *Definite rules should be formulated which will as far as possible control the irrelevant factors in scoring the papers.* The careful use of scoring rules will bring about a definite decrease in the inaccuracy of the pupil scores.

TENTATIVE SCORE CARD FOR RATING ESSAY-TYPE EXAMINATION
QUESTIONS

	Yes	Slightly	No
1. Is the question concerned with important phases of the subject?			
2. If the question emphasizes minor details, are they useful in linking up other facts, ideas, theories, involved in the subject?			
3. Does the question give emphasis to evaluation and to relational thinking?			
4. Is the question apparently of a suitable degree of difficulty in relation to the other questions in the test?			
5. Is the question stated in such a way as to stimulate thought, to challenge interest of pupils?			
6. Does the question force the pupil to integrate his ideas around certain interest-centers?			
7. Is the question stated in such form as to force the pupil to sample widely into his background of fact?			
8. Does the question call for any originality of thought organization and expression?			
9. Does the question call for the pupil to integrate facts gained from different sources?			
10. Is the question limited sufficiently that the pupil has some chance of writing what he really knows about it in a reasonable time?			

The accompanying tentative score card for rating essay-type questions is suggested as a possible means of improving this type of teacher-made examination by calling attention to the desirable qualities in test questions. Unless a question rates "Yes" on the least seven of the ten items, it is certainly of doubtful value and should probably be rewritten or completely eliminated from the examination.

TOPICS FOR DISCUSSION

1. Indicate why there is need for improvement in classroom testing.
2. What are some of the major weaknesses of the oral examination for testing purposes?
3. What uses should the oral quiz be expected to serve in the school?
4. Discuss fully the manner in which limited sampling reduces the reliability of the essay examination
5. List and discuss several factors which contribute to subjectivity of scoring of the typical essay examination
6. Comment upon some of the minor weaknesses of the essay test
7. List and evaluate the advantages which have been attributed to the traditional examination.
8. What are your conclusions concerning the proper place of the essay test in classroom measurement?
9. Identify some of the types of essay questions and indicate key words by which they are introduced
10. Suggest at least five specific devices or procedures for increasing the objectivity of scoring essay-type tests
11. Outline testing procedures by which the essay-type test may be made more effective as a classroom testing technique.

SELECTED REFERENCES

- Caldwell, Otis W., and Courtis, Stuart A., *Then and Now in Education, 1845-1923*, Chapter IV. Yonkers-on-Hudson, N. Y. World Book Co., 1923
- Calkins, Mary Whiton, "Philosophers in Council" *School and Society*, 17 316-20, March 24, 1923
- Cason, Hulsey, "An Intelligence-Question Method of Teaching and Testing." *Pedagogical Seminary and Journal of Genetic Psychology*, 54 359-90, June 1939
- Engelhart, Max D., "Examinations" *Encyclopedia of Educational Research*, pp. 471-78. New York The Macmillan Co., 1941.
- Kandel, I. L., *Examinations and Their Substitutes in the United States*. Bulletin No 28, Carnegie Foundation for the Advancement of Teaching. New York. The Foundation. 1936

- Lang, Albert R, *Modern Methods in Written Examinations*, Chapter IV. Boston : Houghton Mifflin Co, 1930.
- Monroe, Walter S., *Written Examinations and Their Improvement*. University of Illinois Bulletin, Vol. XX, No. 7. Urbana : University of Illinois, 1922.
- Monroe, Walter S, and Carter, R. E., *The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students*. University of Illinois Bulletin, Vol. XX, No. 34. Urbana : University of Illinois, 1923.
- Monroe, Walter S, and Souders, Lloyd B., *The Present Status of Written Examinations and Suggestions for Their Improvement*. University of Illinois Bulletin, Vol. XXI, No. 13. Urbana : University of Illinois, 1923.
- Odell, C W, *Traditional Examinations and New-Type Tests*, Chapter VIII. New York : The Century Co., 1928.
- Odell, Charles W., *The Use of Scales for Rating Pupils' Answers to Thought Questions*. University of Illinois Bulletin, No. 46. Urbana : University of Illinois, 1929.
- Orleans, Jacob S., and Sealy, Glenn A., *Objective Tests*, Chapter II. Yonkers-on-Hudson, N. Y.: World Book Co., 1928.
- Ross, C C, *Measurement in Today's Schools*, Chapter VI. New York : Prentice-Hall, Inc., 1941.
- Sims, Verner M., "Essay Examination Questions Classified on the Basis of Objectivity." *School and Society*, 35:100-2; January 16, 1932.

CHAPTER VIII

CONSTRUCTION AND USE OF INFORMAL OBJECTIVE TESTS

This chapter deals with the following points concerning the construction and classroom use of informal objective tests

- a.* Similarities of the informal objective and standardized tests.
- b.* Major advantages and limitations of the teacher-made objective examination.
- c.* Selecting the content and preparing an informal objective test.
- d.* Administering and scoring the informal objective test.
- e.* Uses and limitations of basic objective item forms.
- f.* Illustrations of objective test item types.
- g.* General suggestions for constructing objective test items.
- h.* Suggestions for constructing basic types of recall and recognition test items.

Introduction. Developments contributing to the improvement of measurements in education have largely followed two main lines: (1) the construction and improvement of standardized tests, and (2) the improvement of teacher-made tests. It is with the second of these that this chapter is concerned. In many respects these two types of measurement are not fundamentally different. Both utilize samplings of subject-matter material to stimulate pupil reactions. In both the performance is expressed in terms of a score. Both make use of exercises which are characterized as being objective. Yet in spite of these similarities, the two types of tests do not seriously overlap in function.

I. CHARACTERISTICS OF CLASSROOM TESTING

The Importance of Classroom Testing. Even though standardized instruments for measuring achievement of school children have come into wide use, the examination set by the teacher still remains the most frequently used means of measuring the achievement of pupils. Although properly constructed standardized educational tests are superior in certain respects to teacher-made examinations,

they will probably never entirely displace them as a means of measuring the results of teaching as indicated by pupil attainment. The teacher frequently has need for a measuring instrument adapted to a particular course of study, or to the emphasis which has been given to the subject in the teaching of a particular class. The informal objective examination, constructed by the teacher to fit the instruction which the class has been receiving, is the obvious answer.

A comparison of the use of the essay-type examination question and the objective (true-false) type of exercise based on the same material is given in the accompanying illustration of an attempt on the part of a teacher to make the test exercises fit a particular unit of instruction. An essay-type question requiring the pupil to organize his thoughts on a unit of civics material is given below along with a pupil's answer to this question which is rated by a number of teachers at about 70 percent of a perfect answer.

Q What are the purposes for which a political party is formed?

A The purposes for which a political party is formed are (1) to establish a platform that will express the views of the members on certain vital questions, (2) to nominate a candidate or candidates for the purpose of running for office. They in turn will uphold the party platform, (3) to separate themselves from other parties with whose principles they do not agree, (4) to establish themselves as a separate organ with certain principles.

The following exercises show how a group of true-false statements based upon the same general subject matter as the foregoing question may be used as an objective-type test. The pupil answered the second of these four exercises incorrectly.

Directions to Pupil Study each statement below. If a statement is True underline True, if False underline False. DO NOT GUESS

- | | | |
|-------------------------------------------------------------------------------------------------------------------|-------------|--------------|
| 1. A political party is formed by a group of people with common political views. | <u>True</u> | False |
| 2. The purpose of a political party is to stimulate interest in political appointments | True | <u>False</u> |
| 3. Political parties usually cater to the opinions of the minority. | True | <u>False</u> |
| 4. The actual accomplishment of political policies is dependent upon the election of those favoring the policies. | True | <u>False</u> |

Standardized vs. Non-Standardized Objective Tests.

Standardized educational tests are structurally not fundamentally different from informal objective examinations in their basic elements. In fact, standardized educational tests are essentially little more than improved and refined objective examinations. Hence, the requirements of test construction suggest the criteria by which the written examination may be improved.

In contrast with their similarities from a structural point of view, the functions of the standardized test and the informal objective examination over the same material are quite distinct. The standardized test, because it is intended for use in many different school systems and in connection with many different types of courses of study, must be general as to content. The maker of a standardized test cannot be sure that its content will actually parallel the instructional emphasis given the subject in the course offered by any individual teacher. Accordingly, the standard test is useful mainly for general comparisons of school with school, class with class, or city with city. It is not designed for use in evaluating the accomplishment of pupils in a class under a particular instructor with a specialized instructional emphasis. By the same reasoning, *the standardized test should probably not be used as the basis for the assignment of class marks in any subject.* The informal objective examination, constructed in accordance with well-recognized principles and incorporating extensive samplings of the subject-matter content actually taught by the teacher, is, on the other hand, a perfectly suitable basis for the assignment of such subject marks. It is quite probable that even though two objective tests, one standardized and one informal, could be made equal in objectivity, length (in terms of number of exercises as well as testing time), and reliability of measurement, their functional values in the classroom would still be quite unlike, because of unavoidable differences in their subject-matter content alone. Thus, in general, standardized and informal objective tests must be considered as having quite distinct and separate functions, and the terms are not to be used interchangeably.

Early writers concerning the teacher-made objective test

differed in their findings and conclusions. Among others, Ruch¹ and Wood² advocated on the basis of experimental studies of the question that objective classroom tests be employed either in conjunction with examinations of the essay type or in place of such examinations. Another experimenter,³ however, concluded that under satisfactory conditions there is little difference in the general merit of the two kinds of measuring instruments. The bulk of the evidence, however, even at that early date in the history of the informal objective test, seemed to favor the liberal use of the objective type of test in all situations in which the measurement of pupil mastery of facts is a major issue.

More recently, Tyler was the leader in a movement to broaden the base for informal objective testing. He pointed out that test content had been validated primarily in terms of the informational content of the courses tested, and recommended a procedure which validated test content in terms of course objectives. Tyler's recommendations for procedures to be followed in achievement test construction are reproduced below without discussion at this point.⁴ Recent enlightened attacks upon construction of both informal objective tests and standardized tests have doubtless been influenced significantly by this point of view.

1. Formulation of course objectives.
2. Definition of each objective in terms of student behavior.
3. Collection of situations in which students will reveal presence or absence of each objective.
4. Presentation of situations to students.
5. Evaluation of student reactions in light of each objective.
6. Determination of objectivity of evaluation.
7. Improvement of objectivity, when necessary.
8. Determination of reliability.

¹ G M Ruch, *The Improvement of the Written Examination*, p 193 Scott, Foresman and Co, Chicago, 1924

² Ben D Wood, *Measurement in Higher Education*, p 337 World Book Co, Yonkers-on-Hudson, N Y, 1923

³ Sterling G Brinkley, *Values of New Types of Examinations in the High School*, p 121 Contributions to Education, No 161 Teachers College, Columbia University, New York, 1924

The results obtained by Brinkley are somewhat at variance with the investigations of almost all others This is possibly because his teachers were inexperienced in the use of new-type examinations and old hands at the use of the traditional examinations

⁴ Ralph W Tyler, "A Generalized Technique for Constructing Achievement Tests" *Educational Research Bulletin*, 10 199-208; April 15, 1931

9. Improvement of reliability, when necessary.
10. Development of more practical methods of measurement, when necessary.

II. ADVANTAGES AND LIMITATIONS OF THE INFORMAL OBJECTIVE EXAMINATION

The foregoing discussion has pointed out that the standardized and the informal objective test are not seriously unlike in the form in which the exercises may be stated. In fact, both types of examinations make use of the same general principles in the formulation of their test exercises. Both the standardized and the non-standardized tests may include enough items to afford consistent measurement. On the other hand, there are a few very distinct differences between the essay-type of examination and the teacher-made objective examination. In general the advantages of the informal objective test are in the areas in which the essay test has definite limitations and perhaps to a less extent the weaknesses of the informal objective test are in areas where the essay test is relatively satisfactory. Therefore, the treatment below is related to that of Section III of Chapter VII and will in some instances depend upon the previous discussion. Because of the similarities between the teacher-made objective test and the standard test noted above, the treatment of this question applies almost equally well to both forms of objective test. Their major differences lie in the purposes for which they are constructed and in the uses to which they are properly put, whereas the discussion here is based more on the form of the types of tests being contrasted.

Advantages of Informal Objective Tests. Of the several merits of the informal objective test, the two most important are answers of the early objective testers to the two major criticisms of the essay examination discussed above—subjectivity of scoring and limited sampling.

Extensive Sampling. Although all tests measure but samples of pupil performance, the objective test by its nature samples so widely that the results obtained from its use closely approximate those which would be obtained if pupil

performance in the subject in question could be measured completely. A test made up of a hundred or so short, well-selected questions or items will adequately sample pupil achievement for many purposes.

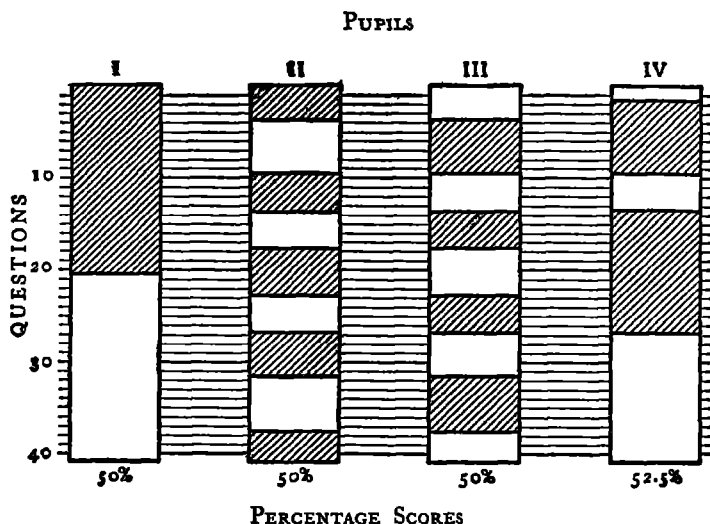


FIGURE 16 THE INFLUENCE OF EXTENSIVE SAMPLING ON TEST SCORES

The results from administering a test consisting of many exercises of narrow range are shown in Figure 16. This illustration is based on the same hypothetical situation as that of Figure 14 on page 134, and the same basic conditions apply. The shaded portions of the four rectangular areas represent the portions of the total course content mastered by Pupils I, II, III, and IV. In each instance the shaded portion is exactly half of the area of the rectangle. The close-ruled horizontal lines represent the 40 short-answer questions which are so distributed as to cover the content of the entire course.

It is apparent here, in contrast with the results shown in Figure 14 when only four questions were used, that the four pupils receive scores which are very similar. Exactly half of the forty lines are opposite the shaded areas for Pupils I, II, and III, so they receive scores of 50 percent.

As twenty-one of the forty marks are opposite the shaded area for Pupil IV, he receives a score of $52\frac{1}{2}$ percent. These results show that the objective test samples widely, and that scores resulting from its use are not likely to be much affected by differences in the specific facts known by different pupils. Enough different questions are asked to make sure that the mark made by each pupil will place him quite accurately in relation to his classmates in terms of his knowledge of course content. This is in direct contrast to the findings based on the illustration of Figure 14, which were that wide differences occurred in the scores assigned to the four pupils.

Objectivity of Scoring. In an objective test the items are so stated that the answers are brief, and usually only one correct answer is possible. A highly objective test may be scored repeatedly by one person with practically no variation in the scores, or it may be scored by a large number of persons with practically no disagreement in the scores assigned. Thus in the new type of examination the exercises may be evaluated on an impersonal basis, entirely independent of the personal judgment of the examiner. This is true, of course, only when the exercises are constructed in accordance with certain recognized principles which are discussed in a later section of this chapter.

Economy of Time. The form in which the objective exercise is stated makes it possible for the pupil to record his response definitely and briefly. This in turn permits many specific reactions to be called for in a relatively brief period of working time. In this way a much wider area of the subject matter may be sampled in a given period, resulting in a higher reliability of measurement per unit of working time.

The brevity and conciseness of the pupil's response make it possible for the scoring of the exercises to be done very accurately and speedily. If the objective examinations are made in accordance with the best practices, the exercises may be scored by simple keys and stencils in the hands of pupils or of ordinary clerical help. The fact that it is possible to score a set of 100-item objective examination papers for a class of thirty-five pupils with a total time expenditure of

not more than three minutes per pupil is a point which even the most conservative teacher will be glad to consider.

Elimination of Bluffing. Fluency of expression and mastery of the language have always been recognized as factors in examinations of the discussion type. Because of the nature of the exercises, the amount of writing done by pupils in responding to objective tests is reduced to a minimum. This practically eliminates bluffing and the advantage which rapid and fluent writers have over those not so gifted. The fact that one pupil can write more material than another in the same length of time should not result in his receiving higher marks in his school subjects.

Possible Disadvantages of Informal Objective Tests. A number of rather important criticisms of objective examinations have been brought forward by teachers and critics. The following list, while not complete, probably contains the more significant of these objections:

Neglect of Training in Organization and Expression of Thought. Teachers sometimes feel that the informal objective test inadequately allows opportunity for the pupil to organize and express his thoughts. One approach to this criticism is through an analysis of how well the essay test fulfills these purposes. Now, it is embarrassingly true that the amount of such training that is derived from the preparation of an essay-type examination paper is almost exactly zero. Literature has not been greatly enriched by samples from examination papers. Few if any prize essays are produced under the stress of writing feverishly for an hour or more on questions or exercises the limits of which are determined mainly by the writers' imaginations. The emotional stress under which the pupil writes his examinations gives him very little opportunity carefully to think his way through the things he actually knows about the subject. He dashes off his thoughts largely in the order in which they come to him. He has almost no time to consider sentence structure, paragraph organization, or vocabulary choice. The net result is that he forms bad rather than good habits of thought, expression, and work.

Some objective testing methods are available for testing the ability of the pupil to organize his thoughts, but no

claim should be made that the objective test provides opportunity for the verbal expression of organized thought. It is probable that the examination should be expected to serve no such purpose, as the opportunity for pupil self-expression in writing can and should be provided adequately elsewhere in the school.

Overemphasis upon Factual Knowledge. This objection to the newer type of objective examination naïvely overlooks the fact that almost uniformly essay-type test questions are based on memory and, in the main, test for the factual aspects of the subject. The thought question as a type is not at all inherent in the essay examination. It is unfortunately almost as rare in the essay-type test as it is in the objective form. Furthermore, there is nothing in the objective form which makes impossible the statement of exercises which stimulate critical and constructive thought. True, most teacher-made tests do not contain this kind of exercise, but that does not mean that they cannot be made when teachers become masters of objective techniques and learn to think deeply enough into the subject matter. The informal objective examination can be used, as is brought out later in this chapter and in subsequent chapters of this volume, in the measurement of various instructional outcomes going in significance far beyond the acquisition of facts and of basic skills. Without doubt the source of this criticism lies not so much in objective methods of measurement in general as it does in the kinds of objective material prepared by the individual teacher.

Encouragement of Guessing. Some teachers and critics believe that there is a tendency for the objective test to encourage guessing to an undue extent. The objective examination form admittedly permits, but does not necessarily encourage, guessing. In fact, it may tend to discourage guessing through its emphasis upon exact knowledges and correct applications and interpretations of factual data, and in its use of correction for guessing formulae in scoring test results. Furthermore, few guesses on objective tests are probably based on pure chance. Rather they are based on slight balances of evidence on one side or the other of an issue on which the pupil is uncertain. Many life activities,

as a matter of fact, are based on chances considerably less than certain of a given outcome. Therefore, it seems that guessing in the sense of weighing available evidence and making the best decision possible is neither injurious to the pupil nor a bad influence upon examination results.

Difficulty of Preparation. The criticism that informal objective tests are difficult to prepare is frequently made. The typical essay test is easy to prepare but hard to score. The informal objective test may be difficult to prepare but it is certainly easy to score. When the advantages accruing to the use of objective tests are balanced against the difficulty of preparing them, the conclusion seems favorable rather than otherwise to the objective test.

Considerable Cost. Experience in the use of objective examinations indicates that they are most valuable when available for classroom use in printed or mimeographed form. Unquestionably the paper cost is an item of expense which in some school systems may be serious. However, some kind of paper must be used for the examination. Mimeograph paper is approximately as cheap as any. If the teacher is willing to do his own mimeographing or hektographing, the extra expense should not be very great. As a matter of actual fact, the cost of preparing objective examinations probably represents one of the very minor items of expense in the average school system, when it is considered in terms of the real educational importance of such equipment. Certain types of examinations appear to be affected very slightly by the practice of dictating the exercises and calling for the objective responses to be recorded on numbered lines on scratch paper. If care is taken in the administration of an objective test in this way, very satisfactory results may be obtained from their oral presentation.

III. CONSTRUCTION AND USE OF INFORMAL OBJECTIVE TESTS

The problems of constructing and using informal objective tests are discussed in this and the following six sections of this chapter. Treated here are the general issues which should receive consideration from the time a test is in the

planning stage to the time when its results have been used finally in the validation of its individual items. The five following sections deal with the various major objective item forms somewhat in detail and present samples of the various item types. Section IX gives general and specific suggestions for drafting items of the five basic types.

The major distinction between recall and recognition forms of items which was presented in Chapter II should be kept in mind. Recall forms are those for which the pupil must depend entirely on himself for furnishing or obtaining the answers. The simple recall and completion items are of this variety. Recognition forms are those for which the material is furnished to the pupil, his responsibility being to decide upon the accuracy of statements, upon the one of several possible answers which is correct, upon relationships existing among lists of items, or otherwise to deal with material presented to him. The basic items of this type are the alternate-response, multiple-choice, and matching.

Content of the Informal Objective Test. It is highly important that the test be definitely based upon the objectives of the course, and also upon the course content. It is true, naturally, that content is basic to a test, and furthermore that the best source of content material is found in the course itself. However, the measurement of factual knowledges and the assumption that the pupil is able to use the knowledges he has acquired, or been modified by, in line with the objectives set up for the course are unsound. Tyler found that knowledge of facts and ability to apply principles to new situations are related only to the degree shown by an average correlation coefficient of not much above .25 in science courses at Ohio State University.⁵ Therefore, not only should the test be so constructed as to measure the degree of attainment of the pupils in the desired outcomes but it should do so by means of test situations which involve the ability to apply and use facts as well as knowledge of facts. Emphasis upon the *why, how, of what significance, with what results, and explain, interpret, give reasons for,*

⁵ Herbert E Hawkes, E F Lindquist, and C R Mann (Editors), *The Construction and Use of Achievement Examinations*, p 7 Houghton Mifflin Co, Boston, 1936

or *give causes* of types of test situations rather than the more highly factual *who*, *what*, *when*, and *why* questions will result in a reduction of weight on the factual aspects of course content.

Care should be taken to sample course content widely and impartially in the selection of materials for a test. It is also ordinarily desirable to use more than one type of objective item in the test, but, on the other hand, not to use too great a variety of item types. For ordinary classroom tests given during one period, two or three types might be used; for longer examinations, variety might be increased by using four or five types or modifications. It should be kept in mind that the subject matter itself is often a factor limiting the types of items used. Recall items place a greater demand upon the pupil's memory of specific facts than is true of recognition items, so it might well be expected that the pupil would recognize the accuracy of certain facts presented to him but not necessarily be able to recall the facts without clues. Therefore, recall items should be used only for important facts.

The test-maker usually finds it advantageous to construct items which fall into large groupings, such as matching exercises, first and then to construct items having narrower scope. It is also desirable to construct multiple-choice items prior to alternate-response forms. This does not mean that all matching and multiple-choice items should be constructed before any true-false or simple recall items are made, but rather that first consideration be given for a certain fact or relationship to the possibility of its use in an item form which is not so flexible and widely applicable as are the true-false and simple recall. If a particular idea does not, for example, readily combine with other similar relationships into a matching exercise and does not furnish enough plausible alternative responses for use in multiple-choice form, it might immediately be set up in one of the simpler forms.

The teacher will find it advantageous, for reasons which will be brought out clearly below, to write each item or each test unit on a filing card or slip. Alternate-response, multiple-choice, and simple recall items should be put on separate cards. Paragraph completion and matching exer-

cises should be written on cards in their entirety, for such test units cannot be broken down by items for listing on separate cards. It is possible and desirable to code these cards in terms of the subject-matter aspects they cover and also to keep records of the use of the item in tests and its validity. More will be said of these last two points in a later section of this chapter.

Assembling and Preparing the Informal Objective Test.

After the test items have been constructed, they should be sorted by types and carefully evaluated in their new settings. There should be a minimum number of items which all pupils can answer correctly or for which no pupils can get the correct answers. A difficulty level averaging about 50 percent is recommended by Lindquist as most satisfactory,⁶ so items should range from that point toward very hard and toward very easy. If there should be too few items of a certain type for a section of a test, those items should be redrafted to fit into one of the sections definitely decided upon.

Test length depends upon many factors other than the nature of the test items and the amount of time available for testing, but these are basic issues to be considered in the preparation of a test. The test should be of such length that all or very nearly all of the pupils can complete it before the end of the testing period. Recommendations have been made concerning the number of items of each type which can be given per unit of time at various age levels. From the available evidence it is impossible to determine in advance the exact working time required for a given form of objective examination. However, a reasonable estimate may be reached by allowing one minute of working time for each two recall items, each three multiple-choice items and each four true-false items. Such recommendations seem to have only very general significance, however, for the difficulty of the items and the age level of the pupils have much to do with time requirements, and teachers vary a great deal in the types of items they construct. The teacher will learn after brief experimentation how long a test should be for a given period of time. The number of items can be

⁶ *Ibid.*, pp. 32-33.

automatically determined by the number which have been constructed when the teacher considers the test to be complete and adequate and of proper length for the testing period. Pupils can advantageously be marked on the test results by a method such as that presented in Chapter XXIII. It is, however, important that a fairly large number of items be used in all objective tests.

Items should be arranged in parts or sections according to type in the final test. Agreement does not exist among test workers concerning the best arrangement of items for informal objective tests. Some prefer arrangement of items in each part by an increasing order of difficulty. If this method is used, the teacher's judgment concerning item difficulty is the only basis for arrangement when items are first used. Item-counting procedures furnish evidence on difficulty after items have been used with a class. Other persons prefer to arrange the items topically within each section of the test, and to consider item difficulty in the arrangement of items only by introducing the test by a few very easy items so that pupils will not become discouraged before they get well started. The authors believe that either organization of the test is satisfactory and that the individual teacher should use the method which better meets the conditions under which he uses the informal objective examination. As arrangement of items by increasing order of difficulty appears to be somewhat the sounder procedure, the accumulation of evidence concerning item difficulty after test items have been used might well warrant the use of that arrangement.

The examination should be prepared for use with the pupils by a mimeographing or other method of reproduction if possible. Some item types can be given orally if necessary, and the blackboard can sometimes be used, at the cost of considerable labor, if sufficient space is available and if it can be kept from the eyes of the pupils until the test is to be given. Complete directions to the pupils should appear on the test folders. This sometimes entails general instructions at the beginning and separate directions for each part of the test. If the item forms are difficult to understand or if pupils are taking objective tests for the first time, samples

showing how they are to record their answers should be given with the directions. The samples should be so simple in content that they will be readily comprehended by all pupils. Numerous illustrations of directions to pupils and of samples to demonstrate methods of answering test items are given later in this chapter.

Pupils should be told in the directions whether or not to guess, and should probably also be told how the test will be scored. The most common procedures and those usually recommended are to instruct the pupils not to guess and then to correct their scores for guessing on alternate-response items. On the other hand, pupils are usually told to attempt each item on the matching test. Frequently the multiple-choice test in which items provide more than three choices is not corrected for guessing.

Administering and Scoring the Informal Objective Test. Little need be said here concerning the administration of the informal objective test except to point out that if the directions to pupils and any necessary sample items are carefully and well prepared the actual administration of the test is simple indeed. The teacher should be careful not to give intentional or unintentional assistance to individual pupils by answering any questions they may ask. The safest procedure is to make certain that the pupils understand how to take the test by careful preparation of the directions, to make sure that individual test items require no explanations by framing them with care, and then to answer no questions about word meanings or interpretations to be placed on certain items while the test is in progress. Pupil questions concerning typographical errors they may encounter in the test should be investigated and the attention of the entire class should be called to any such errors which might within reason cause misinterpretations of items.

Scoring of the test should be by the predetermined method, and should vary with the type of objective item. Scoring keys can be prepared easily by using a copy of the test and cutting it into strip keys and cutout stencils as required. With such keys available, the actual mechanics of scoring the tests are very simple. Each correct answer should ordinarily be given one point of credit. It will be advantageous

to mark each correct answer with a colored pencil for later use for instructional purposes.

Chances of guessing the correct answers vary with different item forms. There is little if any chance of guessing, or at least of making a pure guess, on recall item forms. Obviously the chance is 50-50 on an alternate-response item, but it is only one in five for a multiple-choice item with five alternatives. The correction for chance formula is

$$\text{Score} = \text{Rights} - \frac{\text{Wrongs}}{N-1}, \text{ or } R - \frac{W}{N-1},$$

where N represents the number of possible answers to an item. For the true-false item, this becomes $R - W$. For multiple-choice items of 3, 4, and 5 alternatives, the formula becomes respectively $R - \frac{W}{2}$, $R - \frac{W}{3}$, and $R - \frac{W}{4}$.

Correction for chance is ordinarily used with the true-false test and the multiple-choice test consisting of items which have as few as three alternatives. It need not necessarily be used with multiple-choice items having four or more alternatives, as the chance of making a correct guess is not great in such tests. Matching tests are not corrected for chance, for little opportunity for guessing exists if they are properly constructed.

There should be no attempt to weight individual items of a test differently according to their importance or difficulty. A summary of various studies dealing with this question leads to that conclusion.⁷ It may be desirable in some instances, however, to assign varying weights to the scores resulting from different parts of the test, to account for differences in difficulty or average time required per item, in which case the most satisfactory procedure is probably to multiply by 2 or by 3 the scores from test parts which are thought to be deserving of extra weighting.

⁷ J. Murray Lee and Percival M. Symonds, "New-Type or Objective Tests: A Summary of Recent Investigations (October, 1931-October, 1933)" *Journal of Educational Psychology*, 25 161-84, March 1934.

Anticipating Future Testing Needs. For the teacher who repeats courses annually or more than once each year, concern with a particular informal objective test should not end with the final direct use of the results. Informal objective testing is not economical of teacher time if the teacher starts afresh in the construction of every test over a period of years. Construction of objective classroom tests should be a cumulative and selective process resulting in constant improvement of the tests actually used in the classroom. If tests are to be evaluated and improved in the manner suggested below, test booklets should not be returned to the pupils permanently. However, they may well be distributed for review purposes after the test has been scored and collected when the purpose is accomplished, or used with individual pupils in conferences concerning special points needing further emphasis in their work.

As a means of determining the validities of individual items for future use, the teacher will find the method generally known as item-counting of great value. One of the simple item-counting methods is based on a division of the class into groups of above-average and below-average performance on the test, with about half of the class in each group. The test papers should then be sorted into corresponding groups. The number of correct (or incorrect)⁸ responses to each test item by the pupils in each group can then be determined by a routine clerical procedure. This ordinarily involves the use of squared paper on which the columns represent the items of the test and the rows are used for checking the items correctly (or incorrectly) answered by each pupil. A summation of the check marks in each column for each of the two pupil groups is then made. When the number of correct (or incorrect) responses to each item is converted into a percentage of the number of pupils in the group, data of the type shown in Table III of Chapter V became available.

Such evidence is valuable to the teacher in determining which test items properly discriminate pupil abilities by

⁸ It matters little whether correct or incorrect responses are used, for the two procedures merely result in different methods of showing the same facts.

showing higher percentages of correct (or lower percentages of incorrect) answers for above-average than for below-average pupils, which ones might be suspected of ambiguity or other faults because of failure to effect such discrimination, and which, if any, show reversals of the desired type of discriminative power. If the information concerning item validities thus obtained is recorded on the cards which it was suggested in an above section should be set up for test items and groups of items, the cards become a valuable file for use in the construction of future tests. Items which show the proper type of discrimination can be retained, and those, if any, which discriminate in the wrong direction can be discarded or revised after critical examination reveals the source of their ambiguity or other weakness. The file of cards should include only test items which have been found satisfactory by actual demonstration.

A card file of this type can be used for the construction of new tests when the occasion arises, with assurance that the ambiguous items occurring in previous tests have largely been eliminated. It is possible and desirable to add to the file as course content changes and to withdraw items which, although valid, are no longer applicable because of changing course content and objectives. Need for such constant turnover is greater in the social studies and sciences, in which current developments perhaps have the greatest immediate influence, than in subjects for which the content changes less rapidly, but it is undesirable for any course that objective classroom testing be allowed to become static.

Although this procedure for validating test content may on the surface appear to be lengthy and somewhat involved, the teacher will realize significant dividends in improved pupil measurement by the use of it or some similar procedure. After such a system of keeping a cumulative test item file is once established, the teacher will realize the great saving in time and the increased testing efficiency which results. Time expenditure by the teacher is greatest for the typical essay test in the scoring of pupil results. Time expenditure by the teacher is greatest for the informal objective test in its preparation. Attention to the construction of good tests seems much more defensible than attention to the

scoring of tests which in many instances are not satisfactory measurement instruments.

Practical Uses of Informal Objective Tests. Only brief mention is made here of the uses to which the informal objective examination can be put. The alertness and ingenuity of the teacher largely determine the values which result from his use of the informal objective test.

Informal Objective Tests in Instruction. The evaluation of pupil and class achievement is most effectively accomplished through the use of the objective examination. Even if there were standardized tests for the measurement of most of the outcomes of class instruction, they would be unsuited for this type of use. Properly constructed objective examinations within certain limits aid the teacher in determining points at which instructional adjustments must be made. Pupils, likewise, may be led to discover their specific weaknesses in achievement. Informal objective test results are thus shown to have general diagnostic value for relative pupil strengths and weaknesses. Such tests can also be used for instructional as well as for measurement purposes. Informal objective drill and remedial devices can be constructed by the alert teacher.

Informal Objective Tests in Determining Course Marks. Pupils' scores from valid and reliable objective examinations afford the teacher's best single basis for measuring and rating pupil achievement within a given subject. The results of objective examinations enable the teacher to improve the reliability of his marks if the tests themselves are valid and reliable measures of the course outcomes. Teachers can learn with practice to construct course examinations which will satisfy the criteria of a good examination, and which will be more valid tests for the outcomes of his particular course than standardized tests could ever be. The remaining step for the use of test results in marking is to convert scores to the particular type of marks desired. Because of the importance of this use of informal objective test scores, a widely-used method of converting them to course marks is explained in Chapter XXIII. This system can readily be adapted as required, if it is not applicable in its present form, to the marking system used in a particular school.

IV. SIMPLE RECALL ITEMS

Simple recall test items cannot be definitely distinguished from completion exercises, for the major distinctions appear to rest on complexity and length of the test unit and perhaps on the number of pupil responses it calls for. The simple recall form is by far the most widely used of the recall item types. It usually involves a very brief response by the pupil, such as writing a word, number, symbol, or short phrase in a designated place in answer to a question or to complete a statement.

Uses and Limitations of Simple Recall Items. The simple recall item is best adapted to the measurement of rather highly factual knowledges of the *who*, *what*, *when*, *where* types, and is very widely adaptable to different subject matter in such uses. It can be used to test the ability to identify things described or pictured, in which form it has rather wide range. In identification exercises, it is perhaps best adapted for use with maps and charts in the social studies and representations of biological structures in the natural sciences. It is useful in computational problem situations in arithmetic and the physical sciences.

One of the major characteristics of the simple recall form is its apparent ease of construction, which tends to encourage wider use than is perhaps justified. Because of its tendency to measure factual knowledges rather than understandings, there is danger of overweighting tests with factual materials if the simple recall item is too widely employed. This item is not readily adaptable to the measurement of abilities to apply facts, to perceive complex relationships, and to draw logical inferences. The simple recall form is readily understood by pupils because of its similarity to the essay item.

The simple recall item is not easy to score, because of the tendency for the responses to lack complete objectivity, even though responses may be provided for in terminal and aligned form. It is further limited by the fact that it is not directly adaptable to machine methods of scoring.

Major Types of Simple Recall Items. The simple recall item is perhaps most frequently presented in the form of a declarative statement with a blank in which the pupil is to

write the correct completion occurring at the end of the sentence.

EXCERPT FROM DENNY-NELSON AMERICAN HISTORY TEST⁹

E. DIRECTIONS. On the line in the parentheses at the right, write the word or words required to fill each blank

71. The first negro slaves in America were held in the colony of _____
72. Cortez searched for gold in the country of _____
73. By the Treaty of Paris, in 1763, western Louisiana was given to _____

It also is frequently used, particularly in the lower grades, in the form of a question which is to be answered by the pupil on the line immediately following.

EXCERPT FROM STANFORD ACHIEVEMENT TEST, ARITHMETIC REASONING¹⁰

DIRECTIONS. Find the answers as quickly as you can. Write the answers on the dotted lines.

	Answer
¹ Which is the largest number? 93 67 85 91 89	-----
² How many girls are 3 girls and 6 girls?	-----
³ Mother paid 9 cents for milk and 7 cents for bread. How many cents in all did she pay for these two things?	-----

Another form less widely used but satisfactory involves a list of terms or statements introduced by directions which tell the pupil to write on the line following each the other term or statement called for by the directions.

EXCERPTS FROM MASTER ACHIEVEMENT TESTS, ENGLISH¹¹

Write the **SINGULAR** of each of the following:

- | | |
|------------------|--------------------|
| 1. wharves | 6. halves |
| 2. matches | 7. volcanoes |

⁹ E C Denny and M J Nelson, *Denny-Nelson American History Test* Published by World Book Co., 1928

¹⁰ Truman L Kelley, Giles M Ruch, and Lewis M Terman, *Stanford Achievement Test Arithmetic Reasoning*, Primary Battery Published by World Book Co., 1940

¹¹ *Master Achievement Tests English*, Grade 6 Published by American Education Press.

V. COMPLETION ITEMS

Completion items may be either of the sentence or the paragraph type. Frequently there is little by which a sentence completion item can be distinguished from the simple recall item. The more typical form of the completion exercise, however, is that based on a paragraph of unified material in which several blanks are provided for the pupil to fill with the words, numbers, or short phrases which correctly complete the meaning. Blanks in the completion exercise only occasionally occur at the ends of sentences, so pupil responses typically are scattered over the page. An adaptation of this type of exercise places a number in each blank and similarly numbered blanks at the right-hand margin for use by the pupils in recording their answers. This results in simplifying the scoring procedure for completion exercises.

Uses and Limitations of Completion Items. Similarities between the simple recall item and the sentence and paragraph completion exercise result in considerable similarity of their uses and limitations. Both are typically rather highly factual, but the latter requires the pupil to handle a larger unit of thought and to integrate his ideas more fully. Both are difficult to score objectively, and must be so constructed that the blanks call for definite responses. Neither can be scored directly by mechanical methods. Both may become puzzle situations for the pupil if too much of the thought is omitted from the statement to permit of reasonably quick comprehension of meaning by the pupil. The completion exercise is somewhat harder to score than the simple recall item unless a device which results in aligned and marginal responses is employed.

Completion examples are not so widely adaptable as simple recall items because of the need for broader and more unified thought units in the former. However, the two forms are both useful in a wide variety of subject matter. The completion sentence is applicable, for example, in situations involving use of the correct language form in a given setting in English or the foreign languages, in completing

arithmetical examples of the equation form, and in a variety of situations in the social studies and sciences. The paragraph completion exercise is useful in varied subject matter for situations in which a rather logical chronological, organizational, sequential, or cause and effect type of pattern exists, as, for example, with the processes involved in a complete cycle of blood circulation in the human body.

Major Types of Completion Items. Sentence completion exercises frequently require the filling of two or more blanks by the pupil and the blanks do not, of course, occur at the ends of the sentences, as they typically do in simple recall items.

EXCERPTS FROM UNIT SCALES OF ATTAINMENT IN FOODS AND
HOUSEHOLD MANAGEMENT¹²

Directions for Scale 2: In this Scale there are two kinds of questions. In one kind there are blank spaces to be filled in as in sample A.

- A. The colors of the flag are _____, _____, and _____.
In this you would write **red, white, and blue** in the three blank spaces
1. Strong flavored vegetables should be cooked _____ a cover. 1.
7. A budget is a _____ of future _____. 7.

The paragraph completion exercise differs from the sentence completion mainly by consisting of a longer and perhaps more complex thought unit, probably by requiring more pupil responses, and by consisting of two or more sentences in a well-unified paragraph.

EXCERPT FROM STANFORD ACHIEVEMENT TEST, READING¹³

DIRECTIONS. In the paragraphs below, each number shows where a word has been left out. Read each paragraph carefully, and wherever there is a number decide what word has been left out. Then write the missing word in the answer column at the right, as shown in the sample. Write **JUST ONE WORD** on each line. *Be sure to write each answer on the line that has the same number as the number of the missing word in the paragraph.*

¹² Ethel B Reeve and Clara M Brown, *Unit Scales of Attainment in Foods and Household Management* Published by Educational Test Bureau, 1931

¹³ Truman L Kelley, Giles M Ruch, and Lewis M Terman, *Stanford Achievement Test Reading*, Advanced Battery Published by World Book Co., 1940.

SAMPLE.

Answer

A-B Dick and Tom were playing ball in the field.
Dick was throwing the —A— and —B— was trying
to catch it.

A

ball

B

1-2-3 In olden days men made their own pens
from the quills of feathers. It required consider-
able skill to cut a pen properly so as to suit one's
individual taste in writing. Students were always
on the lookout for good goose, swan, turkey, or other
bird feathers. Goose quills made the most satisfac-
tory —1— for general —2—, but schoolmasters
liked pens made from the —3— of swan feathers
because they fitted best behind the ear.

1.

2.

3

VI. ALTERNATE-RESPONSE ITEMS

Alternate-response items are those in which only two alternatives are presented to the pupil for his response. The simplest and most common forms of alternate-response items are the true-false, requiring an answer concerning the truth or falsity of a statement, and the yes-no, requiring one of those answers to a question. Another form involves the selection of the correct one or better one of two alternatives which are presented as possible completions in a given setting.

The true-false, as the most widely used alternate-response type, has also doubtless been the most popular form of recognition item, and probably remains so today for classroom testing purposes. It typically involves a very simple method of response by the pupil in answer positions in column form at either the left or right side of the test paper.

Uses and Limitations of Alternate-Response Items.

The true-false item is widely applicable in all subject-matter fields. Its ease of construction has resulted in greater popularity and wider use than have been attained by any other item form. However, its ease of construction is frequently delusive, for the elimination of ambiguities from the true-false item is sometimes difficult to accomplish. Lindquist states that this weakness seems to be inherent in the item it-

self, and points out that test technicians are tending to use it less and less.¹⁴ It and the simple recall item are perhaps most frequently taken almost verbatim from textbooks, and consequently in such cases a premium is placed upon photographic memory for facts.

Alternate-response item forms have the advantage of affording coverage of many individual items in a short period of time, since the time requirements are less than for most item types. On the other hand, guessing is more of a problem for this than for any other item type, for which reason little diagnostic value can be obtained by using an item-count method of analyzing the results for a group of pupils or an individual pupil. Alternate-response items are highly objective in scoring, and are readily understood by pupils. This item type is readily scorable by mechanical methods in all of its common varieties.

True-false items can be used satisfactorily in many situations if they are constructed carefully enough to make them largely free from ambiguity. They can be used in testing popular misconceptions and unfounded beliefs in the science and social studies areas. They are also useful for situations in which the absence of enough plausible alternative responses make the use of a multiple-choice item impracticable.

The type of alternate-response form which requires the pupil to select the one of the two alternatives which correctly fills a particular need is very widely useful for measurement of a functional type of instructional outcome in English and the foreign languages. It could be used in a wide variety of situations, but the true-false usually better serves the purpose, so in practice this item form has been limited largely to language usage situations.

Major Types of Alternate-Response Items. The most common form of alternate-response item is the true-false, which may be set up so that the pupil will respond by encircling, underlining, or writing a "T" or an "F," a "+" or

¹⁴ Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann (Editors), *The Construction and Use of Achievement Examinations*, pp. 153-54. Houghton Mifflin Co., Boston, 1936

a “—,” a “+” or a “0,” a “True” or a “False,” or in any one of several other similar ways.

EXCERPTS FROM COOPERATIVE PLANE GEOMETRY TEST¹⁵

Directions: Read these statements and mark each one in the parentheses at the right with a plus sign (+) if you think it is always true, or with a zero (0) if you think it is always or sometimes false.

- | | | | |
|---------------------------------------------------------------------------------------------------------------|------|------------------------------------------------------------------------------------------|-------|
| 1. Any acute angle is greater than its complement. | 1() | 16. A rhombus inscribed in a circle is a square | 16() |
| 2. The lines joining the midpoints of the sides of an equilateral triangle form another equilateral triangle. | 2() | 17. If two angles of a quadrilateral are supplementary, the other two are complementary. | 17() |

Another common form is presented as a question, the pupils' responses usually consisting of encircling, underlining or writing either “Yes” or “No.” This form, which differs little from that presented above, is preferable for use with young children because the situation presented is a very normal one.

EXCERPTS FROM HAGGERTY READING EXAMINATION¹⁶

Draw a line under the right answer to each question.

- | | | |
|----------------------------------------------|-----|----|
| 1. Can good children make promises? | YES | NO |
| 2. Do all people rent houses? | YES | NO |
| 3. Do laborers ever become exhausted? | YES | NO |
| 4. Are compasses used by mariners? | YES | NO |
| 5. Can children act in a serviceable manner? | YES | NO |

An alternate-response form commonly used in English and foreign language tests involves the selection of the proper one of two given word forms for use in a certain setting and indication of the one selected by crossing out the incorrect word form or marking the correct word form.

¹⁵ John A. Long, L. P. Siceloff, and Emma Spaney, *Cooperative Plane Geometry Test*, Form R. Published by Cooperative Test Service, 1941.

¹⁶ M. E. Haggerty and Laura C. Haggerty, *Haggerty Reading Examination*, Sigma 3. Published by World Book Co., 1920.

EXCERPT FROM IOWA EVERY-PUPIL TESTS OF BASIC SKILLS,
LANGUAGE ¹⁷

1. He ☐ did the work himself.
☐ done
2. The cow with the black spots ☐ is eating grass.
☐ are
3. The stick was ☐ broke into two pieces.
☐ broken

VII. MULTIPLE-CHOICE ITEMS

Multiple-choice items have come to be the most popular form for standardized testing of recent years, and are increasingly coming into wide use for informal objective testing as well. A recognition item type, the multiple-choice item commonly consists of an incomplete statement followed by from three to five responses which will complete the statement with varying degrees of accuracy. The pupil is expected to choose the response which correctly completes the statement, and typically to indicate his choice by an answer appearing in a column at the left or the right side of the test paper.

This item type may be in question rather than in statement form or may consist of five words, symbols, or numbers from which the correct one is to be chosen by the pupil. It may request the best of several correct or partially-correct answers on a given point. It may even require responses for the two or more correct answers among those which are furnished.

Uses and Limitations of Multiple-Choice Items. The multiple-choice and its numerous variants perhaps represent the most valuable and at the same time the most widely applicable type of objective test item. It is highly objective in scoring. It is readily, although not necessarily easily, adaptable to the measurement of discriminative power, inferential reasoning, interpretative ability, reasoned understand-

¹⁷ H. F. Spitzer, *Iowa Every-Pupil Tests of Basic Skills: Test C, Basic Language Skills*, Elementary. Published by Houghton Mifflin Co., 1940.

ing, generalizing ability, and other types of outcomes deriving from the pupil's ability to apply and use facts. It is not difficult for pupils to understand and use. It is highly objective, and can be readily scored either by hand or by machine. Item-count procedures based on the results for an individual pupil or a class have considerable diagnostic and analytic significance.

Multiple-choice and multiple-response items in their variety of forms are so widely adaptable to subject matter that the preceding discussion should make the fact evident without illustration. As is the case for the true-false item, there is probably no field of learning to which the multiple-choice item is not widely applicable. However, the necessity for finding at least two and in many cases as many as four plausible responses to go with the correct completion somewhat limits the applicability of the item form within each subject field. Ingenuity on the part of the test-maker and the results of practice in item construction make the item type very widely applicable to the content of various instructional areas, however. Multiple-choice items are not as easily constructed as are some other objective test forms, for there are various technical problems which require great care in the drafting of items. The incorrect answers pupils give to simple recall items often serve as excellent incorrect alternatives if the item is converted to multiple-choice form.

Major Types of Multiple-Choice Items. The basic and probably most common multiple-choice form is that in which the correct completion is to be selected by the pupil from the three to five which are furnished for an incomplete declarative sentence.

EXCERPT FROM CALVERT SCIENCE INFORMATION TEST¹⁸

1. The slipping or movement of the earth's crust is felt as
 ¹ storms ² floods ³ earthquakes ⁴ thunder ()
2. Pressure from inside the earth is sometimes released through
 ¹ rivers ² volcanoes ³ storms ⁴ glaciers ()
3. Since it was first formed the surface of the earth has changed
 ¹ very little ² not at all ()

¹⁸ Everett F Calvert, *Science Information Test*, Intermediate Published by California Test Bureau, 1937

A slight variation of the above is the best-answer form in which two or more of the completions are correct and the pupil is expected to choose the one which best completes the statement.

EXCERPTS FROM COOPERATIVE GENERAL SCIENCE TEST¹⁹

Directions: Each of the following incomplete statements or questions is followed by four or five possible answers. For each item, select the answer that best completes the statement or answers the question, and put its **number** in the parentheses at the right.

- | | |
|--------------------------------------------------------------------------------------------------------|----------------------------------------------|
| 1. Which of the following may be seen in the Northern Hemisphere on a clear night at any time of year? | 7. A man, a dog, a fish, and a bird all have |
| 1-1 Taurus | 7-1 scales. |
| 1-2 Pleiades | 7-2 warm blood. |
| 1-3 Orion | 7-3 lungs. |
| 1-4 Sirius | 7-4 hair. |
| 1-5 The Big Dipper 1() | 7-5 a skeleton 7() |

Another variation, sometimes called the multiple-response, is that in which the pupil is asked to select all of the correct completions from the three to five typically given. There may be only one or as many as several correct answers to different items when this form is used. Each correct response is ordinarily assigned one scoring point of credit. The fact that not only the choice but also the response is plural accounts for the distinction in names between this and the more common multiple-choice item.

EXCERPT FROM INTERMEDIATE SCHOOL AUTO MECHANICS TEST²⁰

Directions for Exercises 75-86 Record your answer to each of the questions in this section as indicated by the question

75. Place a check mark (✓) before three of the following parts that may form part of the electrical system of an automobile
- | | |
|-----------------|-----------------|
| (1) coil | (4) timing gear |
| (2) manifold | (5) commutator |
| (3) distributor | (6) accelerator |

¹⁹ O. E. Underhill and S. R. Powers, *Cooperative General Science Test*, Form Q. Published by Cooperative Test Service, 1940

²⁰ A. D. Althouse, et al, *Intermediate School Auto Mechanics Test*. Published by Department of Instructional Research, Detroit Public Schools.

Somewhat similarly, multiple-choice items can singly or by groups be based on a map, chart, diagram, or table, and require the pupil to interpret the data presented as a basis for answering.

EXCERPT FROM COOPERATIVE MATHEMATICS TEST FOR GRADES
7, 8, and 9²³

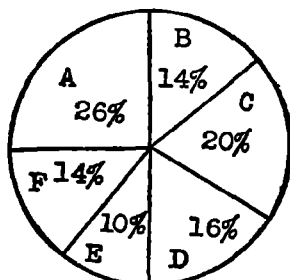


FIGURE 4

31. A pile of 150 blocks was divided into six smaller piles in accordance with the percentages shown in the graph in Figure 4. This indicates that the number of blocks in E was

- | | | |
|------|----|----------|
| 31-1 | 8 | |
| 31-2 | 12 | |
| 31-3 | 15 | |
| 31-4 | 24 | |
| 31-5 | 26 | () |

A multiple-choice form comparable to the type of alternate-response item particularly used in language usage situations employs three or more alternatives.

EXCERPTS FROM MODERN SCHOOL ACHIEVEMENT TESTS, LANGUAGE
USAGE²⁴

- | | | |
|------------------|---------------------------------------|-------|
| | 1. I | |
| 1. My sister and | 2. myself will be glad to see you. | |
| | 3. me | |
| | 1. graduated | |
| 2. He | 2 graduated out of elementary school. | |
| | 3. graduated from | |
| | 1 to | |
| 12. He is | 2. by his grandmother's. | |
| | 3. at | |
| | 1. lay | |
| 13. He | 2. laid his coat on the sofa. | |
| | 3. set | |

²³ *Cooperative Mathematics Test for Grades 7, 8, and 9*, Form RO. Published by Cooperative Test Service, 1941.

²⁴ Arthur I. Gates, et al, *Modern School Achievement Tests, Skill Subjects Language Usage*. Published by Bureau of Publications, Teachers College, Columbia University, 1931.

An illustration of an analogies type of test making use of the multiple-choice response is given below.

EXCERPT FROM ILLINOIS EXAMINATION, ANALOGIES ²⁵

1	eat—bread	drink—water	iron lead stones	1
2	finger—hand	toe—box	foot doll coat	2
3	shoe—foot	hat—kitten	head knife penny.	3
4	dress—women	feathers—bird	neck feet bill.	..	4
5	dog—puppy	. cat—kitten	dog tiger house.	5

VIII. MATCHING EXERCISES

Matching exercises are in effect combinations of multiple-choice items in such manner that the choices are compound in number. Matching exercises differ from all of the objective forms treated previously in the fact that they must occur in groups. There is really no such thing as a matching test item, unless a correct pairing pulled from a group of which it is a part might be so designated. Matching tests are by nature, then, multiple in type, and the number of scoring points is ordinarily determined by the number of responses required of the pupil.

A matching exercise or set usually consists of two lists of related facts between which a constant type of relationship exists throughout. The pupil's responses are expected so to pair items in the two lists as to indicate their proper relationships. Variations involve unbalanced sets, in which more items occur on one side than on the other, sets in which items of one side may be used more than once each, and even compound sets in which double or even triple matchings of all items are necessitated by the provision of three or even four related lists instead of the customary two.

Pupil responses to matching exercises are usually in the form of identifying numbers or letters written in column form in parallel with the items in one of the two or more lists. The unbalanced set has the definite advantage of reducing the chances of guessing the correct answers to practically zero.

²⁵ W. S. Monroe and B. R. Buckingham, *Illinois Examination Analogies*. Published by Public School Publishing Co., 1920.

Uses and Limitations of Matching Exercises. Matching exercises are likely to be rather highly factual in nature, and to make use of the *who*, *what*, *when* and *where* types of relationships and of identifying or naming abilities. They are rather easy to construct, and are perhaps for that reason more widely used than their characteristics warrant. They are likely to include clues to the correct responses unless there is rigid adherence to uniform categories of items in a matching set, and this restriction, desirable though it is, limits at least one side of the test unit to numbers, words, or at least short phrases. This restriction in turn tends to limit use of

EXCERPT FROM SONES-HARRY HIGH SCHOOL ACHIEVEMENT TEST,
MATHEMATICS²⁶

DIRECTIONS. In the parentheses after each geometric condition given below in Column 2 write the number of the result in Column 1 that could be proved by it.

COLUMN 1 (RESULTS)	COLUMN 2 (CONDITIONS)
1. angles equal	66. If two opposite sides are equal and parallel ()66
2. triangles congruent	67. If perpendicular to the same line ()67
3. triangles similar	68. If the sides are proportional ()68
4. lines perpendicular	69. If they have equal arcs ()69
5. lines parallel	70. If side-angle-side equal side-angle-side respectively ()70
6. quadrilateral is a parallelogram	71. If they are parallelograms with equal bases and altitudes ()71
7. parallelogram is a rectangle	72. If their central angles are equal ()72
8. two arcs equal (in same or equal circles)	73. If a tangent is drawn to the radius at point of contact ()73
9. two chords equal (in same or equal circles)	74. If corresponding parts of congruent triangles ()74
10. areas of polygons equivalent	75. If one angle is a right angle ()75

²⁶ W. W. D. Sones and David P. Harry, Jr., *Sones-Harry High School Achievement Test Mathematics* Published by World Book Co., 1929.

the item form mainly to factual types of subject matter.

The matching exercise is economical of space and of construction time. It is useful for matching terms and definitions, names and events, events and dates, books and authors, causes and effects, generalizations and applications, words and symbols, English and foreign words, and many other pairs of related items by use of verbal lists. It is also useful with numbered maps, charts, or pictorial representations for matching places and names, places and events, trends and dates, objects and names, etc., in great variety. The matching exercise appears to be most useful with factual knowledges in a great variety of situations where it is desirable to test over a number of comparable relationships.

Major Types of Matching Exercises. The fundamental form of matching exercise, as shown on the preceding page, has an equal number of items in both lists and involves the use of all of the items in the pairing.

Unbalanced matching sets provide more items on one than on the other side and require that only as many of the items of the longer list be used as have proper pairings with the items of the shorter list.

EXCERPTS FROM COOPERATIVE GENERAL SCIENCE TEST ²⁷

Directions (Items 1 through 12) For each group of items below, place in the parentheses after each word or phrase in the right-hand list the **number** of the word or phrase in the left-hand list with which it is **most directly** associated.

— — — — —		
1 Scarlet fever	4. Rat	4()
2 Diphtheria		
3 Malaria	5. Anopheles mosquito	5()
4 Sleeping sickness		
5 Bubonic plague	6. Tsetse fly	6()
1 Amphibian	10. Amoeba	10()
2 Protozoan (one-celled animal)		
3 Fish	11. Whale	11()
4 Crustacean		
5 Mammal	12. Frog	12()

²⁷ O E Underhill and S. R. Powers, op cit.

Again, it may be that all items in the longer list are to be paired by the use as many times as is individually necessary of the items in the shorter list.

EXCERPT FROM DENNY-NELSON AMERICAN HISTORY TEST²⁸

C. DIRECTIONS Each President's name below is given a number. Show by writing the correct number in the parentheses the administration in which each event occurred.

Washington (1)	46. The sinking of the <i>Lusitania</i>	()
	47. The Dingley Tariff Bill	()
Jackson (2)	48. End of the second United States Bank	()
	49. Invention of the cotton gin	()
	50. Draft riots in New York City	()
Lincoln (3)	51. Establishment of the Federal Reserve Banks	()
	52. The Spanish-American War	()
McKinley (4)	53. The Whisky Insurrection in Pennsylvania	()
	54. The Emancipation Proclamation	()
Wilson (5)	55. The Prohibition Amendment to the Constitution	()

Diagrams, maps, charts, and pictures may be used in what are often called identification exercises by requesting the pupil to match identifying names of places, objects, or parts with their representations in the accompanying figure or picture.

EXCERPT FROM NATIONAL ACHIEVEMENT TESTS, GENERAL SCIENCE²⁹

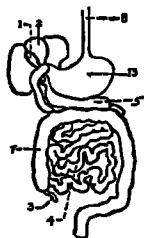


FIGURE 2
Alimentary Canal

3. The small intestine is number_____
4. The stomach is number_____
5. The liver is number_____
6. The pancreas is number_____
7. The gall bladder is number_____
8. The appendix is number_____
9. The large intestine is number_____

²⁸ Denny and Nelson, op cit

²⁹ Robert K. Speer and Samuel Smith, *National Achievement Tests General Science* Published by Acorn Publishing Co., 1939

Additional Examples of Matching Exercises. Listings for the matching exercise need not occur in the common parallel fashion, for one of the item groups may occur above the other without changing the nature of the relationship. The type of setup illustrated below is perhaps most common when the items of one list are to be used anywhere from once to several times each.

EXCERPT FROM METROPOLITAN ACHIEVEMENT TESTS, HISTORY AND CIVICS⁸⁰

Directions. After each event in the list below put the number —

- 1 if it happened before the *Settling of Jamestown*.
- 2 if it happened between the *Settling of Jamestown* and the *Adoption of the Constitution*.
- 3 if it happened between the *Adoption of the Constitution* and the *Civil War*.
- 4 if it happened between the *Civil War* and the *Spanish-American War*.
- 5 if it happened since the *Spanish-American War*.

For example, you should write the number 1 after "Columbus discovered America," because it happened before the *Settling of Jamestown*.

Sample. Columbus discovered America ()

77. Columbus sailed from Europe to find a route to India	()	77
78. The first airplane trip over the Atlantic was made	()	78
79. New York was first settled by the Dutch	()	79
80. Franklin was ambassador to France	()	80
81. Gold was discovered in California	()	81
82. The airplane was invented	()	82
83. Hudson explored the Hudson River	()	83
84. The telephone was invented	()	84
85. The airplane was used for delivery of mail	()	85
86. The Constitutional Convention took place	()	86

A type of recognition exercise which combines matching and completion features requires the pupils to fill blanks in a completion paragraph with dates, names, or terms given in an accompanying list and to indicate their matchings by filling blanks in the paragraph with identifying letters of the dates, names, or terms.

⁸⁰ Richard D. Allen, et al., *Metropolitan Achievement Tests History and Civics*, Intermediate Published by World Book Co., 1933

EXCERPT FROM ELY-KING TESTS IN AMERICAN HISTORY, TEST II ²¹

III. Use the list below to fill the following blanks. Place in the blank the letter that is before the name or date you wish to use:

The Revolutionary War began in the state of . The first real battle of the war was . Independence was declared in the year . The Declaration of Independence was chiefly the work of . The decisive battle of the war was . The last battle of the war was . The commander-in-chief of the American armies was . The treaty of peace which closed the war was signed at .

- | | | |
|---------------|-----------------------------|------------------|
| a. 1789 | h. Legislative | o. Saratoga |
| b. 1775 | i. Congress | p. Massachusetts |
| c. 1776 | j. Senate | q. New York |
| d. 1787 | k. Judicial | r. Paris |
| e. Franklin | l. Executive | s. Philadelphia |
| f. Jefferson | m. Bunker Hill | t. Yorktown |
| g. Washington | n. House of Representatives | |

IX. CONSTRUCTING INFORMAL OBJECTIVE TEST ITEMS

This section of the chapter considers the general principles to be followed in the construction of various objective item types. Such questions as adaptation of item types to various subject matter and the construction and use of the test as a whole were considered earlier in the chapter. Because of the multiplicity of item types, it is impossible to discuss all of them in detail. Therefore, the suggestions are intended mainly for the basic or most common forms of items, although in many instances they are equally well adapted to modified types of the basic items.

General suggestions which seem to be equally applicable to all objective item types are given in the following section. These should serve as the introductory portion of the lists of suggestions on later pages for the various com-

²¹ Lena A. Ely and Edith King, *Ely-King Tests in American History*, Test II. Published by Southern California School Book Depository, 1927

mon or basic forms of items. The student will find that frequent reference to the sample items in the preceding sections will be helpful in the study of methods for constructing the various item types. He is likely to find that some of the samples are not in complete agreement with suggestions made by the writers of this volume, for the samples were chosen to represent various item types broadly rather than to conform in all respects to procedures recommended in this volume. It should be apparent that common sense and personal experience must furnish the basis for recommendations on many issues discussed. Objective evidence is not available concerning the relative merits of different approaches on many of the issues, and on other points only inconclusive evidence and conflicting opinions and practices are presented in the educational literature. Therefore, this section can be said to present the authors' views, based on objective evidence and opinions of others and on their own experience in test construction, on a considerable number of detailed points which must be considered if objective item types are to be well constructed.

General Suggestions for Constructing Objective Items.

A number of suggestions apply equally well to all or most objective item types, so such suggestions are given here. Attention will be given in the subsequent pages to suggestions which apply to recall types and to specific item types of the recognition form.

(1) *Rules governing good language expression should be observed.* This point deserves mention because carelessly-framed and ungrammatical items are more likely to be subject to misinterpretation than are items which are carefully constructed and correctly stated.

(2) *Difficult words should be avoided.* Care should be taken at all times to make certain that the words used in objective items are known to all pupils, for every pupil should be able to understand the intent of all items. This recommendation does not, of course, apply to the technical words of the subject being tested, for knowledge of technical vocabulary is an outcome of instruction which may well be tested. Every effort should be made to adapt general vocabulary words to the ability levels of the pupils being tested,

however. In case of doubt, it is always a safe procedure to choose the simpler of two words which might be used in stating a test item.

(3) *Textbook wording should be avoided.* It is undesirable to obtain items merely by taking a statement from a textbook and using it in its exact textbook form, with a negative inserted, with a word omitted, or otherwise with minor adaptation as a test item. In the first place, an occasional pupil has a memory for specifics of what he has read or heard which would enable him to answer the item in terms of such memory rather than to place responsibility upon his real knowledge and ability. In the second place, a majority of textbook sentences, unless they are from summary paragraphs or are topic sentences of paragraphs, are too detailed to merit direct attention in a test. In the third place, there is danger that items so selected would be too much dependent upon a particular textbook or author and not be broadly representative of the field being tested.

(4) *Ambiguities should be eliminated.* Care should be taken to make certain that each test item is subject to one and only one interpretation. It is not always easy to accomplish this purpose, for ambiguities sometimes remain after an item has been carefully framed and scrutinized. Items should be sufficiently definite that there is no chance for misinterpretation of meaning through reasonable implications or logical inferences. Item-counting methods of evaluating items after they have been used once are helpful in eliminating ambiguities which have been overlooked in the initial framing of a test.

(5) *Items having obvious answers should not be used.* Items to which answers are obvious have no value in a test and should definitely be avoided.

(6) *Clues and suggestions should be avoided.* Items containing clues and suggestions also contribute nothing to a test and may well lack validity.

(7) *Items which can be answered by intelligence alone should not be included.* Items which depend not at all upon knowledge of or ability to apply subject matter but which can be answered by the exercise of intelligent reasoning have no place in an achievement test.

(8) *Quantitative rather than qualitative words should be used.* It is preferable to use words which have quantitative and if possible definite meaning rather than words which are qualitative in nature, as a means of eliminating items which depend upon opinion rather than upon facts.

(9) *Catch words should not be employed.* There is no justification for the inclusion in achievement test items of catch words, misleading statements, or irrelevant confusions. Pupils recognizing such points might interpret such features as typographical errors or as unintentional for other reasons and answer them in terms of what they thought was intended. Furthermore, the best readers, who are frequently the best pupils, are perhaps least likely to note minor errors in a test because rapid reading entails less attention to specific letters and even words than does slow reading.

(10) *Items should not be inter-related.* Items should not ordinarily be so related, at least if they are adjacent or close together in the test, that one depends upon one or more other items in such manner that an answer to the first determines responses for the related items. In effect, such dependence places more than the intended amount of weight upon the first item of any such sequence when answers consistent with the first are given by a pupil for subsequent items.

(11) *Response positions should preferably be aligned.* It is preferable, although not always possible, to have the response positions occur in a columnar arrangement. The pupil is aided by such a consistent position for responses and scoring of the results is greatly facilitated.

Suggestions for Constructing Simple Recall and Completion Items. Several suggestions which are applicable to recall item types alone are given and briefly discussed here. These suggestions represent for recall items a continuation of the list of general suggestions in the preceding pages of this chapter. As the simple recall and completion types are very similar except for two of the following points, the recommendations for these item types are included in one list.

(1) *Lines for responses should be of the same and of adequate length.* In recall item forms the length of all lines or blanks provided for pupil responses should be the same. The lines or blanks should be long enough to provide for

normal writing of the longest word likely to be given as an answer. The constant length of line avoids giving any clue as to the length of the correct answer which might be of use to any pupil in choosing between two answers he might be considering.

(2) *Desired responses should be definite.* Each recall item should require a definite idea or concept as the correct answer in order to reduce the possibility of misunderstanding by the pupil and to insure objectivity of scoring. The response may be a word, a date, a number, a symbol, a formula, an answer to a problem, or even a short phrase.

(3) *Desired responses should be important.* Only important and crucial aspects of a statement should be omitted in recall forms of items, for the omission of secondarily important or unimportant aspects of a statement reduces the significance of the item.

(4) *Any correct answer should receive credit.* Any answer which is correct, whether or not it is the one the teacher expected, should receive credit and the answer should be added to the scoring key for future use.

(5) *Spelling errors probably should not be penalized.* Unless spelling errors occurring in pupil answers are in words technical to the subject for which the test is given, scoring should probably be in terms of the pupil's intent rather than in terms of his spelling accuracy.

(6) *"A" or "an" should not immediately precede a blank.* Either of the indefinite articles restricts the nature of the response word to follow in terms of grammatical correctness, so that the range of possible correct answers is mechanically narrowed for the pupil when "a" or "an" immediately precedes a response position. Employment either of the definite article "the" or of "a(n)," which means either "a" or "an," is permissible, but if possible should be avoided.

(7) *Positions for responses should ordinarily be at the ends of the sentences.* It is perhaps preferable that blanks to be filled occur at the end of sentences rather than in the middle. Statements can usually be so worded that this is easily accomplished.

(8) *Completion paragraphs should be unified wholes.* A completion paragraph should be unified and well organized

and should not consist of several unrelated or poorly related sentences. The pupil's ability to grasp the entire thought unit should be essential to correct responses for the several blanks in the paragraph.

(9) *Completion paragraphs should not obscure the meaning by containing too many blanks.* Sufficient of the paragraph should be given that the meaning is clear to an informed and intelligent reader. It is easy for the teacher constructing a paragraph, who knows definitely what the paragraph is about, to assume unconsciously that the pupil should have the same knowledge and consequently leave out so many words that the meaning is obscure or not ascertainable.

Suggestions for Constructing Alternate-Response Items.

The suggestions below for the alternate-response type of item supplement the general suggestions previously discussed. As the true-false is the most widely-used of these types, most of the suggestions below relate primarily to it or a closely allied form.

(1) *Double negative statements should be avoided.* Double negatives serve no useful purpose, but may cause needless and harmful reading problems for some pupils.

(2) *Statements which are part true and part false should not be used.* Statements should be either true or false, for the use of a true major clause and a false dependent clause or of some other combination of truth and falsity is confusing to the pupil and adds nothing to the test. Although such part true, part false statements are used by some test workers, the result frequently is an unintentional "catch" item.

(3) *"Specific determiners" should be used sparingly and carefully.* Such specific determiners as "always" and "never" occur in false statements much more frequently than in true statements. Statements containing cause or reason clauses also tend to be false more often than true. On the other hand, comparison statements and very long statements are more often true than false.

(4) *Answers should be required in a highly objective form.* It is inadvisable to have pupils write a letter, such as "T" or "F," or a word, such as "True" or "False" in answering the items, for those letters and words look much

alike when poorly written or when written with the attempt to confuse the scorer. Methods requiring pupils to encircle or to underline "T" or "F," "Yes" or "No," "+" or "—," or having pupils mark an "X" in the brackets in either the "T" or "F" column are to be preferred.

(5) *Approximately an equal number of true and false statements should be used.* It is not desirable to have a great disbalance of true and false statements, but on the other hand there is no need for exactly the same number of each type of item.

(6) *Random occurrence of true and false statements should be employed.* A coin may be tossed or some other simple chance procedure be used to make certain that true and false statements will occur in random or chance order.

Suggestions for Constructing Multiple-Choice Items. The following suggestions, supplementing the general recommendations given in an earlier section of this chapter, are primarily for the multiple-choice item type with only one correct answer or the closely-related best-answer type.

(1) *As much of the statement as possible should occur in the introductory portion.* There is no justification for repetition of the same introductory word or words in each of the alternatives; the introductory, or common, portion of the item should include as much as possible as a means of saving space.

(2) *Alternative answers should all be stated in correct grammatical style.* It should be possible to follow the introductory portion of an item with any one of the alternative answers and have the statement be grammatically correct.

(3) *Incorrect alternatives, or confusions, should be plausible.* One or more alternatives which are obviously incorrect in effect give the pupil a greater chance of guessing the correct answer. Pupils' wrong answers to recall items often provide excellent confusions for the same items if put into multiple-choice form.

(4) *"A" or "an" should not ordinarily be used to introduce the alternative answers.* Unless all answers could follow the same article with grammatical correctness, the a(n) device mentioned above or the indefinite article should be used to introduce the alternative answers.

(5) *Items should ordinarily have four or five alternative answers.* Except for use with very young children, four or five alternative answers are preferable as a means of reducing the chances of guessing the correct answer and in order to obtain the desired degree of item difficulty, although two well-chosen confusions are preferable to three or four implausible wrong answers.

(6) *All items should ordinarily have the same number of alternate answers.* Four- and five-response items may be mixed in the same test, although the same number of alternatives for each item is preferable for ease in correction for guessing.

(7) *Alternative answers should ordinarily occur at the end of the statement.* Although the responses may be so placed that additional material common to all is necessary to complete the statement, rewording will ordinarily make possible their placement at the conclusion of the statement.

(8) *Answers should be required in a highly objective form.* It is perhaps preferable that a pupil write the identifying letter or number for the intended response or encircle or otherwise mark it in a special answer column. There is little efficiency in a method requiring underlining or, worse yet, both underlining and otherwise indicating, an intended answer.

(9) *Correct responses should be distributed with approximate equality among possible answer positions.* In four-response items, for example, the first, second, third and fourth alternatives should be correct for approximately the same number of items. It may be desirable to favor the centrally-located responses slightly over first and last responses for the correct answers.

(10) *Random occurrence of correct responses should be employed.* A die may be tossed (disregarding the six) or some other simple chance procedure be used to insure random order in the occurrence of the various correct answer positions.

Suggestions for Constructing Matching Exercises. The suggestions given below for the common type of matching set supplement the general suggestions on pages 188 to 190 for all types of objective items.

(1) *Only one correct matching for each item should be possible.* If items are not mutually exclusive, i.e., subject to only one correct matching, some pupils may be penalized because they happen to choose the one of two or more possible matchings for a certain item which results in the lack of a proper answer for an item at the end of the matching process, when the same number of items appears in each column.

(2) *Consistency of grammatical form should be used.* All items in the left-hand set should agree in form and all items in the right-hand set should likewise be in agreement. It should be possible in so far as the form of the statements is concerned to associate any item of the left with any item of the right column. If this is not true, answers can be obtained partly by attention of the pupil to grammatical detail in the statement of the item.

(3) *Consistency of classifications should be maintained.* Each of the two lists should contain items which are of the same category. Although matching sets which are not consistent within each column are used by some test-makers, the results from mixed categories are sometimes confusing, often provide a means of answering items by the exercise of general intelligence alone, and in general are unsatisfactory. Consistent categories are much to be preferred.

(4) *Matching sets should neither be too long nor too short.* From ten to fifteen pairings are probably optimum for balanced-matching groups. More than fifteen pairs become cumbersome and time consuming. Fewer than ten pairings present opportunities for good guessing on the last few matchings by the pupil who knows most of the pairings. Unbalanced matching sets are definitely preferable if fewer than ten pairs are used, and perhaps should be used in all matching sets.

(5) *Items should be listed in random order in each list.* Such logical arrangements as alphabetical order of first letters of words and chronological order of dates usually accomplish this purpose, for such arrangements are not likely to have any similarity to the relationships between the items of the two lists and furnish no clues to the pupils.

(6) *A set of matching items should always be complete on one page.* The necessity for frequent rereading of items

makes very inefficient any separation of a set of matching items by having it appear on two pages of the test.

(7) *Answers should be required in a highly objective form.* Perhaps the most satisfactory method of providing for pupil responses is to accompany one list with letters or numbers identifying each item and the other list by answer positions, and then to have pupils write the letters or numbers in the answer column in such manner as to indicate their choices.

TOPICS FOR DISCUSSION

1. Explain the differences between standardized tests and informal objective examinations
2. What reasons can you advance for the general conclusion that there is no conflict between standardized tests and informal objective tests?
3. Discuss the major advantages of the informal objective test over the traditional or essay test.
4. Discuss the limitations sometimes claimed for the informal objective test
5. Briefly comment upon the selection of content and general construction of the teacher-made objective test.
6. Discuss pro and con the advisability of using several types of objective test items in the same classroom test.
7. Why should an objective test be difficult enough that no pupil makes a perfect score and yet sufficiently easy that no pupil makes a zero score?
8. What cautions should be observed in administering the teacher-made objective test?
9. How should the various types of objective test items ordinarily be scored?
10. What procedures are useful to the teacher in the revision of the informal objective examination?
11. What are the major uses of the informal objective test?
12. Clearly distinguish between recall and recognition item forms.
13. Distinguish between the two ordinary forms of recall items and illustrate each type.
14. Give examples of several alternate-response item types.
15. Show the differences among the ordinary multiple-choice, the multiple-response, and the best answer item forms. Illustrate.
16. Which type of matching exercise, the balanced or the unbalanced, is preferable? Why?
17. Give some of the most important general suggestions for the construction of objective test items.

18. Supplement the general suggestions for constructing objective test items by giving additional suggestions which apply particularly to the (a) simple recall and completion exercises, (b) alternate-response item, (c) multiple-choice item, and (d) matching exercise.
19. Indicate separately for each of the five objective item types considered in this chapter a few types of facts, skills, or abilities in the measurement of which they are useful.

SELECTED REFERENCES

- Broom, M. E., *Educational Measurements in the Elementary School*, Chapters V-VI. New York McGraw-Hill Book Co, Inc., 1939.
- Brownell, William A., "Some Neglected Criteria for Evaluating Classroom Tests" *Appraising the Elementary School Program* Sixteenth Yearbook of the Department of Elementary School Principals, pp. 485-92. Washington, D. C National Education Association, 1937.
- Engelhart, Max D, "Examinations" *Encyclopedia of Educational Research*, pp. 471-78. New York The Macmillan Co, 1941.
- Gerberich, J. R, "A Technique for Measuring the Ability to Evaluate Objective Test Items." *Journal of Educational Research*, 27 46-50; September 1933.
- Hawkes, Herbert E., Lindquist, E. F., and Mann, C. R (Editors), *The Construction and Use of Achievement Tests*, Chapters II-III. Boston: Houghton Mifflin Co, 1936.
- Kinney, L. B, and Eurich, A. C., "A Summary of Investigations Comparing Different Types of Tests" *School and Society*, 36 540-44, October 22, 1932.
- Lang, Albert R, *Modern Methods in Written Examinations*, Chapter V. Boston Houghton Mifflin Co., 1930.
- Lee, J. Murray, *A Guide to Measurement in Secondary Schools*, Chapters X-XI. New York D Appleton-Century Co, Inc., 1936.
- Lee, J. Murray, and Symonds, Percival M, "New-Type or Objective Tests: A Summary of Recent Investigations." *Journal of Educational Psychology*, 24 21-28, January 1933.
- Lee, J. Murray, and Symonds, Percival M., "New-Type or Objective Tests A Summary of Recent Investigations (October 1931-October 1933)." *Journal of Educational Psychology*, 25:161-84; March 1934.
- Lefever, D. Welty, "Dangers and Values in Teacher-Made Tests." *Education*, 53 409-12, March 1933.
- Lincoln, Edward A., and Workman, Linwood L., *Testing and the Use of Test Results*, Chapter XI. New York The Macmillan Co, 1935.
- Lindquist, E. F., "The Technique of Constructing Tests." *Educational Record*, 15 68-86; January 1934.
- McCall, William A, "A New Kind of School Examination." *Journal of Educational Research*, 1 33-46, January 1920.

- Nelson, M. J., *Tests and Measurements in Elementary Education*, Chapter III. New York The Cordon Co., 1939
- Odell, C W, *Traditional Examinations and New-Type Tests*. New York The Century Co., 1928
- Orleans, Jacob S, and Sealy, Glenn A, *Objective Tests*, Chapters III, V, XIII. Yonkers-on-Hudson, N Y World Book Co., 1928
- Paterson, Donald G., *Preparation and Use of New-Type Examinations*. Yonkers-on-Hudson, N Y World Book Co., 1927.
- Raths, Louis E, "Evaluating the Program of a School." *Educational Research Bulletin*, 17 57-84, March 16, 1938
- Rinsland, Henry D, *Constructing Tests and Grading in Elementary and High School Subjects*. New York Prentice-Hall, Inc., 1938
- Ross, C C, *Measurement in Today's Schools*, Chapter V. New York Prentice-Hall, Inc., 1941
- Ruch, G M, *The Improvement of the Written Examination*. Chicago. Scott, Foresman and Co., 1924.
- Ruch, G M, *The Objective or New-Type Examination*. Chicago Scott, Foresman and Co., 1929.
- Ruch, G M, and Rice, G A, *Specimen Objective Examinations*. Chicago Scott, Foresman and Co., 1930
- Russell, Charles, *Classroom Tests*. Boston Ginn and Co., 1926.
- Tiegs, Ernst W, *Tests and Measurements in the Improvement of Learning*, Chapter IV. Boston Houghton Mifflin Co., 1939.
- Tyler, Ralph W, *Constructing Achievement Tests*. Columbus, Ohio Ohio State University, 1934.
- Tyler, Ralph W, "The Specific Techniques of Investigation Examining and Testing Acquired Knowledge, Skill, and Ability." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XXIX, pp 341-55. Bloomington, Ill. Public School Publishing Co., 1938
- Weidemann, Charles C, *How to Construct the True-False Examination*. Contributions to Education, No. 225. New York Teachers College, Columbia University, 1926.
- Worcester, D A, "Prevalent Errors in New-Type Examinations." *Journal of Educational Research*, 18 48-52; June 1928.

CHAPTER IX

NATURE AND MEASUREMENT OF INTELLIGENCE

The aspects of intelligence and intelligence testing which are given major attention in this chapter are as follows :

- a.* Definitions of intelligence.
- b.* Theories concerning the nature of intelligence.
- c.* Theory of intelligence testing
- d.* Individual tests of general intelligence.
- e.* Group tests of general intelligence.
- f.* Aptitude and readiness tests.
- g.* Performance tests.

This and the following chapter bear a direct relationship to each other. It is important for the student to be conversant with the nature of intelligence and with techniques for its measurement. It is also important that he be able to obtain and use at least the major types of derived scores in furnishing guidance of various types to his pupils. This chapter discusses the theory and measurement of intelligence, while Chapter X goes on from this foundation to present the applied aspects of intelligence and intelligence testing.

Workers in the field of mental abilities are far from agreement both as to the correct terminology to use in discussing mental abilities and also as to the exact nature of the ability or abilities to which the terms apply. It is therefore very difficult to prepare a brief treatment of intelligence and intelligence testing. The discussions of intelligence in this and the following chapter will be based on what the authors believe to be the best modern terminology in this field. The reader will doubtless encounter instances, however, in which test titles and references will not be completely in harmony with the usage to be followed.

The term "mental tests" is here considered in the broad sense mentioned in Chapter III, i. e., as including educational, intelligence, and personality tests. Then those aspects of behavior relating to the intellectual and largely unlearned

abilities of the individual can logically be considered under the heading of intelligence tests. The other two types of mental measurements, educational tests and personality tests, are dealt with elsewhere in this volume.

I. THE NATURE OF INTELLIGENCE

The exact nature of the combination of abilities known as intelligence is not well understood. However, it is definitely known that individuals differ widely in the amount, and perhaps the quality, of it they possess, and that within limits it can be measured.

Definitions of General Intelligence. Many definitions of intelligence have been given. The following list presents some which are most commonly quoted:¹

Colvin. "An individual possesses intelligence in so far as he has learned, or can learn to adjust himself to his environment."

Dearborn. "... the capacity to learn or to profit by experience. . ."

Henmon. "Intelligence . . . involves two factors—the capacity for knowledge and knowledge possessed."

Pintner. "I have always thought of intelligence as the ability of the individual to adapt himself adequately to relatively new situations in life."

Terman. "An individual is intelligent in proportion as he is able to carry on abstract thinking."

Thorndike. "We may . . . define intellect, in general, as the power of good responses from the point of view of truth or fact."

Woodrow. "It is an acquiring-capacity."

Additional definitions taken from Freeman² are:

Binet. "... the tendency of thought to take and maintain a definite direction, the capacity to make adaptations for the purpose of attaining the desired end, and the power of self-criticism."

Burt. "... the power of readjustment to relatively novel situations. . ."

Stern. "... the general mental adaptability to new problems and conditions of life."

The above definitions seem to fall into at least three patterns—the rather formal definitions stressing mainly what

¹ Symposium, "Intelligence and Its Measurement" *Journal of Educational Psychology*, 12 123-47, 195-216, March and April 1921.

² Frank N. Freeman, *Mental Tests Their History, Principles, and Applications* (Revised Edition), p. 248. Houghton Mifflin Co., Boston, 1939.

have been called the higher mental powers, the definitions emphasizing ability to learn, and the definitions placing major emphasis upon adaptability. It is felt that the last type of definition particularly, by which intelligence is conceived as *the ability of the individual to adapt himself* to his environment and to new situations, is the most meaningful for the purposes of the teacher. However, the fact that ability to learn and ability to think in abstract terms are both evidences of intelligence should not be overlooked.

Freeman³ lists three concepts of intelligence—the organic, the social, and the psychological or behavioristic. He considers that the third is the only one which is of direct concern to intelligence testers and calls the others factors in intelligence. The psychological or behavioristic concept accepts as intelligence the types of behavior which are measured by intelligence tests. Intelligence has been defined as “that which intelligence tests measure.” This definition is in line with Freeman’s psychological or behavioristic concept. The definition has meaning, for it implies that intelligence, although it has not yet been adequately defined or delimited, conditions the individual’s behavior and that it is, therefore, through observation and measurement of his behavior that his intelligence can be estimated.

Theories Concerning Intelligence. Theories concerning the nature of ability go back as far as pronouncements of the early philosophers. However, only three of the most important theories of the last century are presented here. Two of them are important to the user of intelligence tests because of the manner in which they have modified and are now modifying testing practices.

The Faculty Theory. According to the faculty theory, intelligence consists of a number of relatively independent and largely correlated and specialized abilities of various types, such as memory, imagination, honesty, and language ability, to name only a few. The closely related theory of formal discipline maintained that these faculties could be developed individually by means of general mental exercise.

³ Frank N. Freeman, “The Meaning of Intelligence” *Intelligence Its Nature and Nurture* Thirty-Ninth Yearbook of the National Society for the Study of Education, Part I, Chapter I, pp. 11-20. Public School Publishing Co., Bloomington, Ill., 1940.

However, when the theory of formal discipline was disproved and the transfer of training concept directed attention to the fact that such faculties as those named above are neither psychological entities nor subject to general training, the faculty theory was forced into the discard as an explanation of mental abilities.

The Two-Factor Theory. Spearman first presented his two-factor theory in 1904.⁴ He proposed a general factor, or *g*, which enters into all types of performance, and many specific factors, called *s*, which combine with *g* to determine total activity. Basing his theory on technical statistical relationships and treatments of data, Spearman later added a third type of factor, called *group factors*, which represent the overlap among *s* factors.⁵ Thus, according to his theory, a *g* or general factor which might be called energy, *group factors*, such as number ability and mechanical ability; and many *s* or specific factors constitute ability.

Primary Mental Abilities. Spearman's work may be considered the forerunner of the present *factor analysis* approach to the nature of mental ability. Among the factor analysts is Thurstone, who has isolated the seven factors of perceptual, number, verbal, spatial, memory, inductive reasoning, and deductive reasoning,⁶ which he calls primary mental abilities. These primary abilities might appear on the surface to relate closely to the "faculties" of the early psychologies, but the factors emerging from the work of Thurstone and other factor analysts not only are substantiated by correlational relationships but also appear to have sound psychological evidence to support their existence.

II. THE MEASUREMENT OF INTELLIGENCE

Indirect Measurement of Intelligence. For practical purposes, intelligence has been defined in a preceding sec-

⁴ C. Spearman, "General Intelligence Objectively Determined and Measured" *American Journal of Psychology*, 15 201-93, 1904.

⁵ C. Spearman, *The Abilities of Man*, p. 82 The Macmillan Co., New York, 1927.

⁶ Louis L. Thurstone, *Primary Mental Abilities* Psychometric Monograph Series, No. 1 University of Chicago Press, Chicago, 1938.

tion of this chapter as the ability to learn or to adapt to new situations. These definitions imply that this type of ability is subject to evaluation in a rather direct manner. Such is not the case, however, for ability to learn can only be inferred from the fact that learning has occurred in a test situation. Since intelligence itself cannot be measured, test-makers can only measure the performance of tasks the successful completion of which is generally believed to be dependent upon intelligence. The value of the intelligence test lies in the fact that it affords an objective basis for this inference. It samples widely from the fields of learning resulting from experiences assumed to be common to all persons subjected to the test. The pupil's capacity to learn is determined by summing up his reactions to the items of the test. No intelligence test measures capacity directly, but within limits such tests reflect the individual's ability to learn by measuring his ability to react to fragments of his environment. The quality of this reaction is evaluated generally in terms of the average reactions of large groups of unselected individuals. That is, most intelligence tests consist of batteries of different types of tests sampling into many different fields of interest and ability.

There is apparently no way of determining very precisely which particular fields of human interest or ability should be sampled in the attempt to secure this cross-section of mental activity. On this account there are many different intelligence tests and few of the authors of these tests agree with one another as to the proper content of such measuring instruments. It is important that the sampling be sufficiently diverse and representative to permit the securing of an estimate in the nature of an average which will not penalize a person because he may not have had this or that specific experience. Briefly, the measure or average obtained from a test which does sample representative reactions is taken to be truly indicative of one's ability to learn. Roughly it is assumed that what an individual has done with his mind by way of learning is indicative of the kind of mind he has. Differences in intelligence test scores are probably sufficiently accurate, rough as they are, to indicate such differences in

mental power. Back of all this is the generally accepted idea that the extent to which one has learned through experience is proportionate to his capacity to learn.

Factual Content of Intelligence Tests. It has been contended, and not without justification, that intelligence tests do not differ from achievement tests, inasmuch as both are founded upon the measurement of knowledges and skills which have largely been learned. Intelligence is sometimes defined in terms of what a person has learned, on the theory that in a normal environment a person's learning is commensurate with his ability to learn.

Obviously a test of ability to learn must have some type of content. Intelligence tests admittedly contain factual materials. Such tests attempt to measure abilities to see relationships, to draw reasoned inferences, to manipulate, to compare, to contrast, and otherwise to handle factual materials which themselves are so commonly known and at such low difficulty levels that all persons who have had any but the most exceptional environment backgrounds should know the necessary facts and have the necessary skills for understanding and taking, although not necessarily for succeeding upon, the tests. To contend that intelligence tests have been completely successful in eliminating the significance of the factual content would be foolhardy and contrary to available evidence.

A few intelligence tests contain vocabulary sections requiring considerable knowledge of word meanings for successful performance. Several tests also measure facts in general in widely-studied areas of knowledge. The justification for the inclusion of such factual items in an intelligence test is that opportunities are supposed to be similar for all persons experiencing a normal environment to learn such facts and that the degree to which different persons do so is partial evidence concerning their intellectual levels. More frequently, however, intelligence tests attempt, but with varying degrees of success, to rule out or at least to minimize the influence of environment upon an individual's test performance and thereby to obtain a measure of his innate or unlearned abilities.

Kelley⁷ has stated that general intelligence tests and achievement tests overlap to the degree indicated by a correlation coefficient of .90. In general, coefficients of .40 to .60 are found between tested intelligence and academic achievement, but higher degrees of relationship are sometimes found. When such correlations approach .70 or .80, the intelligence test is looked upon with suspicion by some and may be considered a general scholastic achievement test rather than an intelligence test.⁸

Cattell, believing that general intelligence tests measure acquired knowledges and skills to a considerable degree and also that they frequently test abilities of too specific a nature, devised a culture-free test.⁹ The test items, largely pictorial rather than verbal, were chosen to measure abilities to run pencil mazes, to build up series, to classify, and to determine relationships of varying degrees of complexity. The content was so selected as to be independent in large degree of acquired or learned *meaning*, so that he thinks the test can be given with equal fairness to persons reared in any civilized society and even, by pantomime, to primitive peoples.

The teacher should probably admit that intelligence tests in varying degrees test factual knowledges which not all pupils have had equal opportunities to learn, but he is probably justified in the belief that such knowledges are at a minimum in at least the better tests and that the environments of pupils in the typical school are sufficiently similar that all pupils have had approximately equal opportunities to learn such facts as are included in the tests.

III. TYPES OF GENERAL INTELLIGENCE TESTS

General intelligence tests, both individual and group types, are discussed and illustrated below, so that the student may obtain a more complete understanding of the characteristics

⁷ Truman L. Kelley, *Interpretation of Educational Measurements*, p. 208. World Book Co., Yonkers-on-Hudson, N. Y., 1927

⁸ Paul L. Boynton, "Intelligence" *Encyclopedia of Educational Research*, p. 630. The Macmillan Co., New York, 1941

⁹ Raymond B. Cattell, "A Culture-Free Intelligence Test I" *Journal of Educational Psychology*, 31 161-79, March 1940

and representative content of these important instruments for the measurement of general mental ability.

General Intelligence Scales—Individual. Individual intelligence examinations constitute the most accurate devices for the measurement of intelligence. The length of the test, the wide variety of reactions called for, the fact that the subject receives his instructions personally from the examiner, the fact that the examiner is afforded an opportunity to observe each reaction made by the subject, and the careful standardization of procedures for administering the test and scoring the subject's reactions all contribute to the high degree of accuracy. The full time of an examiner is required for each pupil tested. The examiner must be a person who is more capable and efficient in test administration than is the average teacher. Furthermore, he must be one who has had training and much experience in giving individual intelligence tests.

Individual intelligence tests are largely patterned upon the *Binet-Simon* tests brought out in France from 1905 to 1911. The first Binet test, published in 1905, was followed by the Binet-Simon revisions of 1908 and 1911. American adaptations and revisions were published by Goddard in 1911, Kuhlmann in 1912, Terman in 1916, Herring in 1922, and Terman and Merrill in 1937. The Terman and Merrill *New Revised Stanford-Binet Tests of Intelligence* is today, as was its 1916 predecessor until recently, the best known and most widely used individual test of general intelligence in America.

The general procedure in administering the *New Stanford-Binet* is representative of that of many of the other revisions mentioned. The type of performance varies considerably with the different exercises. These exercises are presented to the child by means of spoken directions. The examiner tests one child at a time. The test should be given in a quiet room where there is freedom from distraction. A friendly attitude between examiner and subject should be maintained. The examiner is directed to make sure that the subject understands what is to be done, and in all cases the burden of proof is with the examiner to show that the subject has responded in a way that is representative of his

ability. The following quotation¹⁰ emphasizes the importance of establishing rapport with the pupil to be examined :

The examiner's first task is to win the confidence of the child and overcome any timidity he may feel in the presence of a stranger. Unless rapport has been established, the results of the first tests are likely to be misleading. The time and effort necessary for accomplishing this are variable factors, depending upon the personality of both the examiner and the subject. It is impossible to give specific rules for the guidance of the examiner in establishing rapport. . . The examiner must himself be genuinely interested and friendly or no amount of skilled technique will enable him to establish a sympathetic, understanding relationship with children.

After rapport has been established, the examiner starts to test the child with materials at a scale level on which he is likely to succeed with some effort. If he is successful on all tests at this level, the examiner, assuming that he could pass all tests at lower levels, passes on to the higher levels and continues on through the scale until the subject fails all tests at one age level. In effect, the child has been tested over the entire scale, for his success on all tests at one age level makes almost certain that he could pass all tests at lower levels and his failure on all tests of another, and higher, age level indicates with essential certainty that he could go no higher on the scale.¹¹

The child's mental age is determined by giving him credit for the number of years below the level on which he passes all tests and adding to this amount the years and months of credit assigned to the higher-level tests he passes. For reliable use of the *Stanford-Binet* test, the examiner must be well versed in the standardized procedures for administering the test and for scoring pupil responses, as well as possess the essential ability to establish rapport with the child. Ordinarily from thirty minutes, for young children, to one hour, for older persons, is required for giving the test.

It is not feasible here to reproduce more than a few sample test elements, but the two following samples from the *New Stanford-Binet*, chosen from those most easy to reproduce in limited space, will give the student some idea of the nature of the test.

¹⁰ Lewis M. Terman and Maud A. Merrill, *Measuring Intelligence*, pp. 46-57. Houghton Mifflin Co., Boston, 1937.

¹¹ *Ibid.* p. 63.

YEAR III-6, FORM L, TEST 3, COMPARISON OF STICKS¹²

Comparison of Sticks

Material: Match sticks, cut to 2-inch and 2½-inch lengths.

Procedure: Place the two sticks on the table before the child in the positions indicated below and about an inch apart. Say, "Which stick is longer?" "Put your finger on the long one." Give three trials, alternating the relative positions of the long and the short stick. In case one of the first three trials is failed, give three additional trials, continuing to alternate the positions of the sticks.

(a) _____ (b) _____ (c) _____

Score: 3 of 3 or 5 of 6.

SUPERIOR ADULT I, FORM L, TEST 2, ENCLOSED BOX PROBLEM¹³

Enclosed Box Problem

Material: Any small cardboard box.

Procedure: Show S. a box and say:

(a) "Listen carefully. Let's suppose that this box has 2 smaller boxes inside it, and each one of the smaller boxes contains a little tiny box. How many boxes are there altogether, counting the big one?"

(b) "Now let's suppose that this box has 2 smaller boxes inside it, and that each of the smaller boxes contains 2 tiny boxes. How many boxes are there altogether?"

(c) "Now suppose that this box has 3 smaller boxes inside it and that each of the smaller boxes contains 3 tiny boxes. How many boxes are there altogether?"

(d) "Now suppose that this box has 4 smaller boxes inside it and that each of the smaller boxes contains 4 tiny boxes. How many are there altogether?"

Score: 3 plus.

The lists of test titles¹⁴ at several age levels of the Form L *Stanford-Binet* between Year II and the Superior Adult III, which represent the bottom and top of the scale, will indicate the variety of abilities tested, the scalar arrangement of tests from easy to difficult, and the duplication at different age

¹² Ibid. p. 84.

¹³ Ibid. p. 125.

¹⁴ Ibid. pp. 75-132.

levels of similar types of test situations at varying levels of difficulty.

Year II

Three-Hole Form Board
Identifying Objects by Name
Identifying Parts of the Body
Block Building · Tower
Picture Vocabulary
Word Combinations

Year V

Picture Completion · Man
Paper Folding · Triangle
Definitions
Copying a Square
Memory for Sentences II
Counting Four Objects

Year VIII

Vocabulary
Memory for Stories · The Wet Fall
Verbal Absurdities I
Similarities and Differences
Comprehension IV
Memory for Sentences III

Year XII

Vocabulary
Verbal Absurdities II
Response to Pictures II
Repeating 5 Digits Reversed
Abstract Words II
Minkus Completion

Average Adult

Vocabulary
Codes
Differences between Abstract Words
Arithmetical Reasoning
Proverbs I
Ingenuity
Memory for Sentences V
Reconciliation of Opposites

Superior Adult III

Vocabulary
Orientation Direction II
Opposite Analogies II
Paper Cutting II
Reasoning
Repeating 9 Digits

General Intelligence Tests—Group. Group intelligence tests originated in America during the First World War. The *Army Alpha* and *Army Beta* tests, the latter really a performance scale, were developed for use in selecting Army recruits for officers' training and for other positions requiring high intelligence. Shortly after the War, Otis, Terman, and others began to bring out group tests devised for use in the schools, and many such tests were published between 1918 and 1925. Although new tests have made their appearance from time to time since 1925, the last few years have been more productive in this respect than was the period from 1925 to 1935.

Space limitations prevent the use of illustrations from more than a few intelligence tests and permit only a brief treatment of the testing techniques used. No attempt is made to furnish descriptions of any of the group tests of general intelligence. Instead, sample items of various types representative of testing techniques are shown and briefly commented upon. The only way by which the student can become truly familiar with intelligence tests is by examination and actual use of them.

The accompanying illustration of two of the *Kuhlmann-Anderson Intelligence Tests* shows parts which measure knowledge of the alphabet and ability to follow directions of various types. These are among the higher-level tests of a series of 39 which are divided into nine booklets for use from the first grade to maturity. Bases are provided for interpreting the results in terms of the intelligence quotient (IQ) and also in terms of the percent of average development (PA), both of which are discussed in Chapter X.

The *California Tests of Mental Maturity* are illustrated by samples from the Pretest, and from one of the non-language and one of the language parts in Test 2 and Test 5 respectively for Grades 7 to 10. The pretest, unique among intelligence tests so far as the authors know, is not a part of the test proper but is designed for the purpose of locating those pupils who cannot in fairness to themselves take the remainder of the test because of inadequate vision. Mental ages and intelligence quotients can be obtained separately for the language and non-language sections and for the

EXCERPTS FROM KUHLMANN-ANDERSON INTELLIGENCE TESTS,
TESTS 27 AND 35¹⁵

TEST 27

1. The fifth letter of the alphabet is 1
2. The second letter before the last letter is 2
3. The third letter before M is 3
4. The letter midway between H and N is 4
5. The second letter after the fourth letter is 5
6. The letter two letters to the right of the letter E is 6
7. The first letter to the left of the tenth letter is 7
8. The letters of the word the in the order in which
they come in the alphabet are 8
9. The letters of the word boy in the order in which
they come are 9
10. The word you get by putting the first letter
between the two middle letters of the alphabet is 10

TEST 35

Draw a line through the middle letter in the longer of these two words: Revenge, Assert Write here a word of five letters meaning the opposite of *slow*. Write here

a word which rhymes with *hay* and means a part of a week.

Draw a line after each of these two letters A B making the first line half as long as the second. Think what year this is, then write here the digits in the reverse order, the one which belongs last coming first Cross out one digit in each of these numbers which does not appear in the other number: 43689, 64378

instrument as a whole. These are among the few tests which provide not only a measure of general mental ability but also measures based on parts which are devoted to some of the major factors of mental ability.

¹⁵ F Kuhlmann and Rose G Anderson, *Kuhlmann-Anderson Intelligence Tests*, Fifth Edition Published by Educational Test Bureau, 1940.

EXCERPTS FROM CALIFORNIA SHORT-FORM TEST OF MENTAL MATURITY¹⁰





















PRETEST A.

Directions In each group of letters and numbers, put a circle around the letters and numbers in the second row that are the same as those in the first row of the group.

		D	E	L	3	
A.	V	Ⓓ	O	Ⓛ	C	Ⓢ
		H	O	X	5	
I	A	Z	N	H	X	S
	
6	P	Q	B	V	K	H

TEST 2

Directions In each row find a drawing that is either the same drawing or different views of the first drawing. Put an X on the line under this drawing and put the number of the drawing you mark on the line to the right.

    	    
    	    
1 _____ 2 _____ 3 _____ 4 _____ 5 _____	1 _____ 2 _____ 3 _____ 4 _____ 5 _____

TEST 5

Directions Read each group of statements and draw a line under the correct logical answer. Write the number of this answer on the line to the right.

- 0 All four footed creatures are animals
All horses are four footed. Therefore
1 Creatures other than horses can walk
2 All horses can walk
3 All horses are animals

3 0

- 1 Mr. X is an aviator.
Mr. X is scout master for his home town. Therefore
1 Aviators make good scout masters
2 One aviator is a scout master
3 Scout masters make good aviators

1

- 8 Either your sister is more intelligent than you, or as intelligent or less intelligent.
But your sister is not more intelligent, nor is she less intelligent. Therefore

- 1 Your sister is less intelligent than you
2 Your sister is as intelligent as you
3 Your sister is more intelligent than you

3

- 9 Jim has a better batting average than Ed.
Ed has a better batting average than Bill.
Which has the best batting average?

- 1 Jim 2 Bill 3 Ed

3

Another test which furnishes two part scores and a total score is the *American Council on Education Psychological Examination for High School Students*. Samples from the completion and the number series tests, respectively from the

¹⁰ Elizabeth F. Sullivan, Willis W. Clark, and Ernest W. Ties, *California Short-Form Tests of Mental Maturity*, Intermediate S-Form. Published by California Test Bureau, 1939.

EXCERPTS FROM AMERICAN COUNCIL ON EDUCATION PSYCHOLOGICAL
EXAMINATION FOR HIGH SCHOOL STUDENTS¹⁷

COMPLETION

Think of the missing word in each sentence below.
Then mark the first letter of that word.

A (4) is a tract of land devoted to agricultural purposes.

C :...: D :...: F :...: G :...: H :...:

An (3) is a long, slender wooden implement for propelling or steering a boat

A :...: E :...: I :...: O :...: U :...:

A (7) is a song to quiet babies

D :...: F :...: G :...: K :...: L :...:

A (3) is a representation of the earth's surface or a part of it.

K . M . N :...: P :...: R :...:

A (12) is a box or room for keeping food cool.

D :...: F :...: N :...: Q :...: R :...:

NUMBER SERIES

In each series below, find the rule and mark the next number.

9	13	17	21	25	29	33	34	35	37	40	41
							:...:	:...:	:...:	:...:	:...:
82	73	64	55	46	37	28	14	18	19	20	27
							:...:	:...:	:...:	:...:	:...:
14	19	24	29	34	39	44	48	49	50	54	59
							:...:	:...:	:...:	:...:	:...:
17	19	16	18	15	17	14	11	12	13	15	16
							:...:	:...:	:...:	:...:	:...:
2	4	12	14	42	44	132	133	134	260	268	396

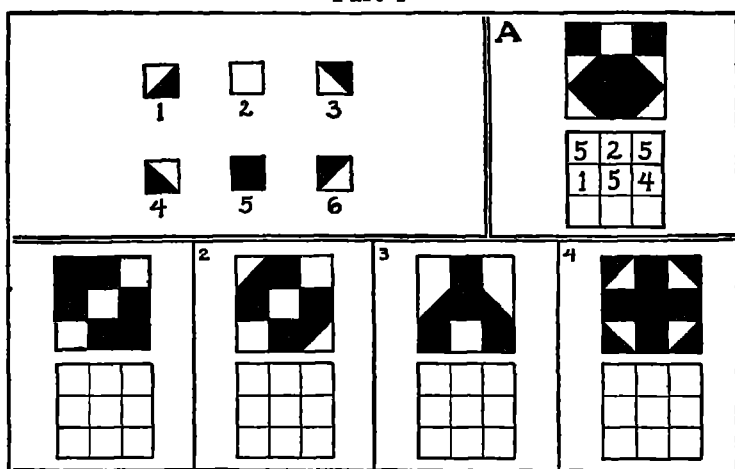
linguistic and quantitative sections, are given in the accompanying illustration. Although the publishers do not claim that the part scores represent primary mental abilities, they

¹⁷ L. L. Thurstone and Thelma Gwinn Thurstone, *American Council on Education Psychological Examination for High School Students*, 1910 Edition Published by American Council on Education.

say that they represent two "groups of abilities significant for curricula that are dominantly linguistic or technical" and that "there seems to be justification for using the two principal subscores as well as the total or gross score"¹⁸ for counseling purposes. The test, published in a new edition annually, is provided with percentile norms for the total and two part scores which permit direct comparison of results from year to year.

EXCERPT FROM DETROIT ALPHA INTELLIGENCE TEST¹⁹

Part 4



The *Detroit Alpha Intelligence Test*, for Grades 5 to 9, is illustrated by the accompanying sample from Part 4, in which the pupil is to identify each of the nine squares composing the larger squares in terms of the patterns of shading by writing into each square the key number for the particular pattern of shading. Sample A and oral directions given by the examiner inform the pupils of the method they are to follow in taking this part. Norms are furnished in such

¹⁸ *Manual of Instructions American Council on Education Psychological Examination for High School Students*, 1939 Edition, p. 2 American Council on Education, Washington, D C

¹⁹ Harry J Baker, *Detroit Alpha Intelligence Test* Published by Public School Publishing Co, 1924

form that either letter ratings or intelligence quotients can be used in interpreting the results.

IV. TYPES OF SPECIFIC INTELLIGENCE TESTS

Aptitude Tests. Specific intelligence testing dates back to 1913, when Münsterberg tested telephone girls and street-car motormen for speed, observation, memory, attention, and accuracy. During the last fifteen years or so, aptitude tests have appeared for a number of areas of performance, such as those involved in various occupations in the trades and industry, various broad areas of performance commonly dealt with in the school, and various narrow areas of performance largely unique to the school. The various types of aptitude tests largely possess in common the characteristic of testing the individual's potentialities in terms of the specific abilities resulting from inheritance and general experience but of disregarding the abilities resulting from specific training or education. Thus, aptitude tests parallel intelligence tests, although they are narrower in scope. They are often called specific intelligence tests.

Teachers and school officers, aside from those engaged in vocational guidance and placement, are more concerned with aptitudes for school subjects and fields of study than with occupational areas. Therefore, occupational aptitude tests, sometimes called trade tests, will not be discussed intensively in this volume, but will receive treatment only insofar as some of them are useful in the schools.

Among the first tests of aptitude to be developed primarily for school use were several for mechanical, musical, artistic, and clerical abilities. In the academic areas of English, foreign languages, mathematics, and the sciences, the *Iowa Placement Examinations, Aptitude Series*, published in 1925, appear to be the pioneer instruments. These tests, primarily useful at the college level, were followed by other aptitude tests for algebra and geometry, English, the foreign languages, mathematics, and the sciences for secondary school use.

The variety of areas of behavior served by aptitude tests makes impracticable a comprehensive discussion of such in-

struments here. They will receive consideration in Chapters XIV to XX by subject fields, parallel with prognostic tests, which, although frequently measuring the results of training, have somewhat similar uses. Aside from tests in the music and art fields, aptitude tests are devised almost exclusively for use at the high school and college levels.

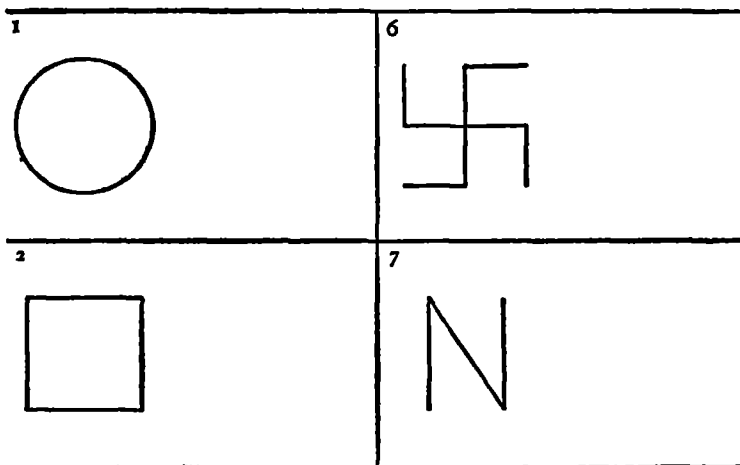
One aptitude test, the *Detroit General Aptitude Examination*, merits mention here, however, because it combines in the one instrument measures of general intelligence, mechanical aptitude, and clerical aptitude, and furnishes separate norms for the general and two specific aspects of ability as well as for several combinations of test parts which have interpretative significance.

Readiness Tests. Readiness tests, found primarily in reading and arithmetic, are largely tests of specific intelligence, in that they measure the results of inheritance and general training rather than of direct instruction. As readiness tests imply by their general designation, they measure readiness to undertake a new type of activity which is dependent upon the maturation of various physical and mental abilities. They may in one sense be considered as aptitude tests at the elementary and even the primary school levels, where they almost entirely occur. Although some of the reading and arithmetic tests published prior to 1930 may indirectly have served the same purpose, readiness tests as such seem to have been published since that date. Recognition of the values of pre-testing in those subjects to determine what ones of the pupils are ready for instruction seems also to have developed somewhat parallel to the publication of the tests.

Tests of this type are usually restricted in applicability to a particular subject field. However, the *Metropolitan Readiness Tests* are devised for determining the readiness of a child to learn first-grade skills of all types, and consequently are briefly discussed and illustrated here. The six parts of the test seem to measure the types of abilities used primarily in reading, number work, and handwriting. Tests 2, 4, and 5, for which the instructions are given orally by the examiner and which require few skills in pencil manipulation of any complexity, measure respectively ability to copy

EXCERPTS FROM METROPOLITAN READINESS TEST²⁰

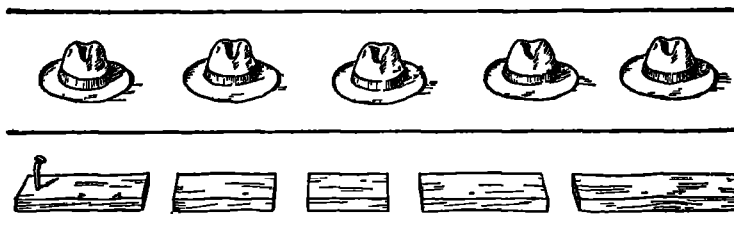
TEST 2. COPYING



TEST 4. SENTENCES



TEST 5. NUMBERS



²⁰ Gertrude H. Hildreth and Nellie L. Griffiths, *Metropolitan Readiness Tests*. Published by World Book Co., 1933

simple figures and forms, sentence comprehension, and comprehension of number and quantitative relationships.

V. TYPES OF PERFORMANCE TESTS

Performance tests require motor or manual rather than verbal responses. In their simplest form, language is required neither in administering the tests nor in responding to them. With the exception of certain form board tests for measuring complex types of mechanical aptitude, they are devised mainly for use with very young children, with mental defectives, and with persons unable to use English with reasonable efficiency. Therefore, their primary purpose seems to be the measurement of abilities not requiring language proficiency, or the measurement of abilities in certain types of persons for whom tests demanding reading and writing are precluded by their language handicaps. Both illiterates and persons who can read, write, and speak a foreign language with fluency but who are deficient in the ability to use English are included in this last group.

Two types of performance tests may be distinguished—those requiring the use of a pencil for marking, but not for writing, and those requiring manipulations of various items of testing equipment.

The *Army Beta* test for use with adults who cannot read, write, or perhaps even understand English illustrates the first type. Directions are given by pantomime, and the subjects respond by tracing mazes, indicating whether groups of numbers are alike or unlike, supplying missing elements in pictures, etc. The accompanying illustration shows a few sample items from various parts of the *Kellogg-Morton Revised Beta Examination*. Results from this test can be interpreted in terms of mental ages.

The second type of test, requiring manipulation of apparatus, depends largely upon form boards which are not unlike jig-saw puzzles. The accompanying reproduction of the tests comprising the *Pintner-Paterson "Long" Performance Scale* shows the general nature of form boards used in the measurement of mental ability. Directions are usually given orally by the examiner. The pupil's success is measured by

EXCERPTS FROM KELLOGG-MORTON REVISED BETA
EXAMINATION²¹

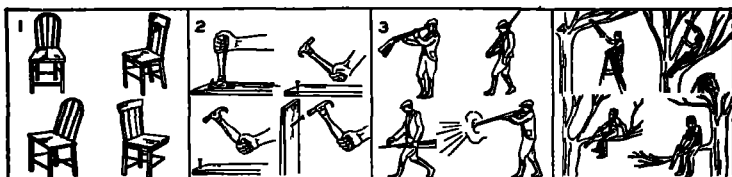
TEST 1

Mark the shortest path from each arrow at the left to the opposite
arrow at the right, but do not cross any of the lines



TEST 2

In each square mark the thing that is wrong



TEST 3

In each picture draw what is left out. Work fast



TEST 4

Look at each pair of drawings or numbers, and make a mark on the dotted line
if they are not alike



time, errors, moves, and other evidences of success or failure. Fifteen separate tests, each of which nets a mental age score, are included in the scale. The median of these mental ages is taken as the pupil's mental ability measure.

²¹ C. E. Kellogg and N. W. Morton, *Revised Beta Examination*. Published by The Psychological Corporation, 1935.

9. Discuss the theoretical foundation upon which specific intelligence tests depend
10. Distinguish between aptitude tests and readiness tests.
11. For what purposes are performance tests ordinarily used?
12. Indicate the nature of performance tests.

SELECTED REFERENCES

- Anastasi, Anne, *Differential Psychology*, Chapter XI. New York: The Macmillan Co., 1937
- Bingham, Walter V., *Aptitudes and Aptitude Testing*, Chapter IV. New York: Harper and Brothers, 1937.
- Boynton, Paul L., "Intelligence and Intelligence Tests." *Encyclopedia of Educational Research*, pp. 622-34. New York: The Macmillan Co., 1941
- Broom, M. E., *Educational Measurements in the Elementary School*, Chapters XI-XII. New York: McGraw-Hill Book Co., Inc., 1939.
- Buros, Oscar Krisen, (Editor), *The 1938 Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938.
- Buros, Oscar Krisen, (Editor), *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: The Mental Measurements Yearbook, 1941.
- Freeman, Frank N., "The Meaning of Intelligence." *Intelligence. Its Nature and Nurture*. Thirty-Ninth Yearbook of the National Society for the Study of Education, Part I, Chapter I, pp. 11-20. Bloomington, Ill.: Public School Publishing Co., 1940.
- Freeman, Frank N., *Mental Tests: Their History, Principles, and Applications* (Revised Edition). Boston: Houghton Mifflin Co., 1939
- Garrett, Henry E., and Schneck, Matthew R., *Psychological Tests, Methods, and Results*. New York: Harper and Brothers, 1933.
- Gilliland, A. R., Jordan, R. H., and Freeman, Frank S., *Educational Measurements and the Class-Room Teacher* (Revised Edition), Chapter XVII. New York: The Century Co., 1931.
- Goodenough, Florence L., *Measurement of Intelligence by Drawings*. Yonkers-on-Hudson, N. Y.: World Book Co., 1926.
- Holzinger, Karl J., "Factor Analysis." *Encyclopedia of Educational Research*, pp. 487-92. New York: The Macmillan Co., 1941.
- Hull, Clark, *Aptitude Testing*. Yonkers-on-Hudson, N. Y.: World Book Co., 1928.
- Hunt, Thelma, *Measurement in Psychology*, Parts II-III, Chapter VII. New York: Prentice-Hall, Inc., 1936.
- Levine, Albert J., and Marks, Louis, *Testing Intelligence and Achievement*, Chapters III-V. New York: The Macmillan Co., 1928.
- Madsen, I. N., *Educational Measurement in the Elementary Grades*, Chapters V-VI. Yonkers-on-Hudson, N. Y.: World Book Co., 1930.

- Nelson, M. J., *Tests and Measurements in Elementary Education*, Chapter XI. New York The Cordon Co., 1939.
- Odell, C. W., *Educational Measurements in High School*, Chapter XV. New York The Century Co., 1930
- Orleans, Jacob S., *Measurement in Education*, Chapter 3. New York: Thomas Nelson and Sons, 1937
- Peterson, Joseph, *Early Conceptions and Tests of Intelligence* Yonkers-on-Hudson, N Y World Book Co., 1925.
- Pintner, Rudolf, *Intelligence Testing* (New Edition). New York: Henry Holt and Co., 1931.
- Sandiford, Peter, *Foundations of Educational Psychology*, Chapter V. New York Longmans, Green and Co., 1939
- Smith, Henry L., and Wight, Wendell W., *Tests and Measurements*, Chapter XIX. New York Silver, Burdett and Co., 1928.
- Spearman, Charles, *The Abilities of Man*. New York The Macmillan Co., 1927.
- Stoddard, George D., "Contributions to Education of Scientific Knowledge about Mental Growth and Development" *The Scientific Movement in Education* Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XXXIV, pp. 421-34. Bloomington, Ill. Public School Publishing Co., 1938
- Stoddard, George D. (Chairman), *Intelligence: Its Nature and Nurture* Thirty-Ninth Yearbook of the National Society for the Study of Education, Parts I and II. Bloomington, Ill. Public School Publishing Co., 1940
- Symonds, Percival M., *Measurement in Secondary Education*, Chapter IV New York The Macmillan Co., 1928
- Symposium, "Intelligence and Its Measurement" *Journal of Educational Psychology*, 12 123-47, 195-216, March and April 1921
- Terman, Lewis M., and Merrill, Maud A., *Measuring Intelligence*. Boston Houghton Mifflin Co., 1937
- Thurstone, Louis L., *Primary Mental Abilities* Psychometric Monograph Series, No 1. Chicago University of Chicago Press, 1938.
- Watson, Goodwin, "The Specific Techniques of Investigation Testing Intelligence, Aptitudes, and Personality" *The Scientific Movement in Education* Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XXX, pp. 357-73. Bloomington, Ill Public School Publishing Co., 1938
- Webb, L. W., and Shotwell, Anna Markt, *Testing in the Elementary School*, Chapters V-VII. New York: Farrar and Rinehart, Inc., 1939.

CHAPTER X

USING INTELLIGENCE TESTS IN PUPIL GUIDANCE

[This chapter presents a discussion of the following points concerning the use of intelligence test results:

- a.* Procedures in intelligence testing.
- b.* Scores derived from intelligence tests.
- c.* Distribution of intelligence.
- d.* Classroom uses of intelligence test results.
- e.* Relationship between intelligence and achievement.

A large part of the rapidly growing popularity of intelligence tests among teachers and supervisors may be traced to three main causes: (1) the tests themselves have been greatly improved in the accuracy and value of the resulting measures; (2) a larger proportion of school officers have become intimately acquainted with intelligence tests and testing procedure with a correspondingly greater appreciation of the functions they serve; and (3) the changes in modern conceptions of education and attitudes toward it have made the utilization of such devices almost essential. Therefore, such devices are important tools of enlightened teaching procedure.

I. GENERAL PROCEDURES FOR INTELLIGENCE TESTING

Administering and Scoring Intelligence Tests. During the early years of the intelligence testing movement, the classroom teacher was given little part in the testing procedures and frequently was even denied access to the results. However, as teachers have become more conversant with intelligence testing techniques and the use of results, they have gradually been given more responsibility in the administration and scoring of the tests and in the interpretation and use of results of group intelligence tests. As has been pointed out in the preceding chapter, the administration of individual intelligence tests should still remain a responsi-

bility of the clinical psychologist rather than that of the classroom teacher.

In many schools today, teachers administer, score, and quite often interpret the results of group intelligence tests. However, too great care cannot be taken by the teacher who participates in an intelligence testing program to understand the procedures for administering and scoring the tests and to follow the practices recommended by the test author, for it is only by such strict adherence to proper methods that reliability of the results is assured.

Care in the Use of Intelligence Test Results. On the whole, intelligence tests seem secure in the place they now hold as indispensable supporting tools for achievement tests, as valuable instruments for the more exact classification of pupils, and as guides to the teacher in matters of pupil behavior and conduct and to the pupil himself in certain vocational and related matters. There are, however, a few dangers attached to their careless or indiscriminate use which the teacher and administrator should guard against. The more important of these dangers are probably social in their character. In the first place, there is the danger which may arise through giving publicity to the results of intelligence testing. There is probably no other place in the use of educational measurements where so much real "social dynamite" is to be found. The parents of children are reasonably tolerant of intelligence testing devices until their own children are measured and make low scores. If their children are tested in arithmetic and show up poorly, it is a matter of small consequence, for that is the teacher's fault and reflects on the school alone. But let one of their own children show up poorly on an intelligence test and it is a different matter. They believe that mental ability is inherited, and the results then do reflect on the parents. In the long run, nothing but damage will ordinarily be done by using intelligence test results for any other than purely school purposes, and then they should be used in strict confidence.

A second danger in the careless use of such tests lies in the effect which knowledge of his own intelligence may have on the individual. The fact that a child is informed that he is unusually brilliant may release a flood of egotism

on his part, or, depending upon his personality, subject him to the gibes of his fellow pupils. The fact that a child is publicly labeled as slow mentally may place a stigma on him which can never be removed. The safest practice is to restrict knowledge concerning results of intelligence testing to responsible school officers and teachers in the main, to make such information available to parents only in occasional and well-considered instances where need arises, and to withhold such information from pupils themselves until they reach at least senior high school or, preferably, the college level. In no case does it seem justifiable to make intelligence quotients of individual pupils known to any persons other than their teachers and school officers, their parents, and themselves.

II. DERIVED RESULTS OF INTELLIGENCE TESTING

A raw score from a test has little or no meaning unless it can be compared in some manner with other similarly obtained and comparable raw scores. This general principle applies to intelligence tests as well as to achievement tests. Therefore, it is important that the teacher shall know the meaning of, and the method of obtaining, the most common types of derived measures used in the interpretation of intelligence test results. As the methods of obtaining the most important of the derived scores discussed below are given fully in Chapter XXIII, only general meanings are discussed below.

Mental Age (MA). Terman defines mental age as "that degree of general mental ability which is possessed by the average child of corresponding chronological age," and as "an index of absolute mental level" indicating "the level of development which a child has reached at a given time."¹ For example, a child has a mental age of ten years if his level of mental development is equal to that of the normal child of exactly ten years. Thus, if a representative group of pupils, all of whom are ten years of age, makes an average score of 45 on an intelligence test which is being standardized,

¹ Lewis M. Terman, *The Intelligence of School Children*, pp. 7-8 Houghton Mifflin Co., Boston, 1916.

any pupil who subsequently takes this test and earns a point score of 45 is said to have a mental age of ten years. An average score for each age group is established in the same manner.

The mental age (MA) is a measure of *mental level* or of *mental maturity* of the individual. Taken alone, it tells nothing of how relatively bright or dull the child may be, but it does give an indication of the level of ability at which the child potentially can work. For example, information to the effect that a certain child has a mental age of 7-6 does not enable a person to judge whether the child is bright, average, or dull. It is only when he knows or at least can estimate the child's chronological age that he can draw conclusions concerning the child's brightness.

The mental age should probably be considered a specific rather than a general concept. That is, a child does not have just one mental age at a given time; he has many.² His mental age, then, depends upon the particular test or tests by which it has been determined, and such tests may be specific intelligence or even personality tests as well as general intelligence tests, although the latter are the only tests in the areas of mental ability commonly providing mental age norms.

Intelligence Quotient (IQ). When the chronological age (CA), i.e., life age in years and months, is known for a pupil, and his mental age (MA) has been determined from his point score on an intelligence test, his intelligence quotient (IQ) can be computed. The intelligence quotient is a simple method of expressing the relationship between a pupil's mental age and his chronological age. To obtain the IQ, a child's mental age (in months) is divided by his chronological age (in months), the result is multiplied by 100, and the whole number nearest to the result is taken as his intelligence quotient. The formula is.

$$IQ = 100 \frac{MA}{CA}.$$

² Lewis M. Terman and Maud A. Merrill, *Measuring Intelligence*, p. 25 Houghton Mifflin Co., Boston, 1937

If this formula is applied for a child who has a mental age of twelve years six months (12-6) when he is ten years five months (10-5) of age chronologically, the following is the result :

$$IQ = 100 \frac{12-6}{10-5} = 100 \frac{150}{125} = 120.$$

The intelligence quotient is a measure of the pupil's relative *brightness*. If it is assumed that an average child grows in mentality at the same rate as he ages chronologically, it then appears that children who have IQs over 100 are above average and children who have IQs below 100 are below average. This is not in disharmony with the usual indication of normal intelligence as being represented by IQs between 90 and 110, for people of normal intelligence center around but are not necessarily exactly at the average of intelligence. However, as this concept of the average is applicable only in terms of the population as a whole and as very few pupil groups are average in this sense, the teacher should not generalize this statement and make it apply to pupil groups in the school. The IQ alone tells nothing about the level of work of which a child is capable, for two children of age six and age twelve might both have IQs of 110 and yet the younger child would be entirely incapable at that time of types of performance commonplace to the older child.

The Mental Growth Curve. The curve of mental growth has long been under scrutiny and has been subjected directly and indirectly to many research studies by psychologists. However, no completely satisfactory unit of mental growth has yet been found. This fact, which results from the lack of an absolute zero point of intelligence, from the lack of a simple and constant mental growth unit, and other technical reasons, gives rise to a major problem in the measurement of intelligence for persons beyond their late-middle teens in chronological age. In practice, intelligence tests handle this problem in various ways, but a usual method is to use the individual's actual chronological age in computing the IQ until he attains the age of fourteen to eighteen and from that point to assume for purposes of computing his intelli-

gence quotient that his chronological age remains constant for the remainder of his life. The justification for doing so is found in the shape of the mental growth curve. Progressing upward very rapidly during early life, and slowing down somewhat during childhood and the early teens, it flattens out to almost a horizontal line by the age of sixteen or so. Although Thorndike has presented evidence to show that mental growth continues into the early twenties,³ the annual increments or additions beyond the age of sixteen are very small indeed.

Constancy of the IQ. A heated controversy over the constancy of the intelligence quotient has been waged during the last few years. Although it has been recognized for many years that the IQ obtained by the use of the best modern tests fluctuates within limits because the tests are not perfect, and that major environmental changes for an individual may well be reflected in his IQ, rather startling evidence was presented some years ago⁴ to show average gains of twenty IQ points for 600 children who had attended pre-school for four years. Later and more startling evidence⁵ showed that children of dull parentage who were placed in foster homes shortly after birth had mean intelligence quotients of 116 when they were tested a few years later. These and other studies support the belief that the intelligence quotient is significantly influenced by very favorable environments.

Although such findings have not been uniformly obtained by experimenters,⁶ they are supported by other types of experimental evidence revealing at least the possibility of marked changes in intelligence quotients as the result of improved environments.⁷ Stoddard sums up the case for

³ Edward L. Thorndike, Elsie O. Bregman, and Ella Woodyard, *Adult Learning*, p. 127. The Macmillan Co., New York, 1928.

⁴ Beth L. Wellman, "The Effect of Pre-School Attendance on the IQ" *Journal of Experimental Education*, 1:48-69, September 1932.

⁵ Harold M. Skeels, "Mental Development of Children in Foster Homes" *Journal of Consulting Psychology*, 2:33-43, March-April 1938.

⁶ Florence L. Goodenough and Katharine M. Maurer, "The Mental Development of Nursery-School Children Compared with that of Non-Nursery-School Children" *Intelligence Its Nature and Nurture* Thirty-Ninth Yearbook of the National Society for the Study of Education, Part II, Chapter IX, pp. 161-78. Public School Publishing Co., Bloomington, Ill., 1940.

⁷ Percival M. Symonds, "Psychological Tests and Their Uses: Review and Preview" *Review of Educational Research*, 8:217-20, June 1938.

inconstancy of the IQ⁸ and points out Binet's expression of the belief⁹ that the IQ is subject to improvement under desirable conditions of stimulation.

The answer to this question may never be known for certain. In fact, as is brought out later in this chapter, the IQ itself is under attack and may in time be replaced by a more satisfactory measure. However, the vast majority of school children do not undergo such radical changes of environment during their school careers that the problem is of great practical significance to the teacher. Yet, there are questions concerning motivation, emotional adjustment, optimum placement of pupils, and many others which bear significantly upon pupil performances not only on intelligence tests but also on achievement tests and in scholarship, so the teacher should at least be aware of this controversial issue and some of its implications.

Future of the IQ. It is apparent from the above discussion that the intelligence quotient is far from a perfect measure of brightness. It appears to be a more accurate measure for the years of middle childhood than for the first years of life or post-adolescent years. Its constancy seems to be strongly in question. These weaknesses and others of a more technical nature raise logical questions concerning its continued and final acceptance as the best measure of brightness, although it is still one of the most satisfactory measures from which to predict success in school and is highly useful in pupil guidance. The alternative methods discussed below for indicating intelligence represent attempts to obtain a more satisfactory measure.

Freeman, after analyzing the problem carefully, states that :

It may be true that the IQ is more convenient, but it is a question whether its inherent ambiguity does not make it better policy to adopt the statistically superior standard score and to educate teachers to understand and use it.¹⁰

⁸ George D Stoddard, "The IQ Its Ups and Downs" *Educational Record*, 20 44-57, Supplement No 12, January 1939

⁹ Alfred Binet, *Les Idees Modernes sur les Enfants*, p 146 Ernest Flammarion, Paris, 1909.

¹⁰ Frank N Freeman, *Mental Tests Their History, Principles, and Applications* (Revised Edition), p. 105. Houghton Mifflin Co, Boston, 1939.

Another attack on the IQ¹¹ recommends that the age scale method of measuring intelligence be abolished, advocates the replacement of the mental age concept by a combination of measures from separate tests, and takes the stand that the controversy concerning the constancy of the IQ is largely futile because its constancy or inconstancy does not depend upon fundamental issues but upon the manner in which tests provide means of obtaining the IQ.

Although the teacher should certainly understand the nature and proper uses of the IQ, he should also have some realization of its limitations, technical though they may be, and should be alert to the alternative methods for designating levels of intelligence which have been developed and which may be evolved in the future. The presentation of several alternatives below should take on additional significance in view of the apparent waning of prestige of the IQ.

Personal Constant (PC). Heinis developed the personal constant¹² for the purpose of obtaining a measure which would be more accurate than the IQ for persons of very superior and very inferior intelligence levels. The measure, which he now prefers to call the *percent of average development*, but which is better known as the *personal constant*, is intended to give quantitative expression to the normal curve of mental growth in terms of growth units which have constant meaning at all age levels. The PC is computed by converting both the mental age and the chronological age to growth units by the use of a table of mental growth units,¹³ dividing the MA value by the CA value, and multiplying by 100. Thus, the PC involves the substitution of growth units for MA and CA in the IQ formula.

Although Kuhlmann recommends that users of the *Kuhlmann-Anderson Intelligence Tests* employ it rather than the IQ¹⁴ and Hilden finds that the PC fluctuates less

¹¹ M. W. Richardson, "The Logic of Age Scales" *Educational and Psychological Measurement*, 1 25-34, January 1941

¹² H. Heinis, "A Personal Constant" *Journal of Educational Psychology*, 17 163-86, March 1926

¹³ A. H. Hilden, *Table of Percent of Average Development Based on Mental Growth Units* Educational Test Bureau, Minneapolis, 1936

¹⁴ F. Kuhlmann and Rose G. Anderson, *Instruction Manual Kuhlmann-Anderson Intelligence Tests*, Fifth Edition, p. 17 Educational Test Bureau, Minneapolis, 1940

than the IQ,¹⁵ Cattell finds the IQ to be definitely more constant for bright children and somewhat less constant for dull children than is the PC.¹⁶ Freeman notes that the computation of the personal constant is more time consuming than is that of the intelligence quotient, and indicates that the evidence now available concerning the values of the PC is inconclusive.¹⁷

Index of Brightness (IB). The index of brightness is stated in the same form as the IQ. While its meaning is somewhat similar to that of the IQ, it is derived in quite a different manner. In this case, the pupil's relative brightness is expressed as a positive or negative deviation from the norm of pupils of his age. The difference between a pupil's score and the norm for persons of the same chronological age is added to (if his score is above the norm) or subtracted from (if his score is below the norm) 100 to obtain his index of brightness. Otis, who uses the measure for his *Quick-Scoring Group Tests of Mental Ability*, himself states that the index of brightness has the same significance as an intelligence quotient.¹⁸ Freeman, however, points out that the method by which the IB is derived makes improbable its consistency with the IQ.¹⁹

Percentile Scores. Percentile scores (also called centile scores) are frequently used to indicate a pupil's status in intelligence. This method is used particularly at the high school and college levels, for the intelligence quotient, as has been pointed out above, is not as meaningful a measure for post-adolescent and adult years as it is for periods of childhood and adolescence. The percentile score describes a pupil's placement in an age or grade group in terms of the percentage of the group scoring lower than he does. The *American Council on Education Psychological Examinations* at both the high school and college levels present norms for

¹⁵ Arnold H. Hilden, "A Comparative Study of the Intelligence Quotient and Heims' Personal Constant" *Journal of Applied Psychology*, 17 355-75, August 1933.

¹⁶ Psyche Cattell, "The Heims Personal Constant as a Substitute for the IQ." *Journal of Educational Psychology*, 24 221-28, March 1933.

¹⁷ Freeman, op cit p 296.

¹⁸ Arthur S. Otis, *Manual of Directions for Gamma Test Otis Quick-Scoring Mental Ability Tests*, p 4. World Book Co., Yonkers-on-Hudson, N Y, 1937.

¹⁹ Freeman, op cit p. 300.

the interpretation of scores in terms of percentiles for different grade levels.

Standard Scores. Another type of measure which indicates a pupil's intelligence level in terms of his position within a certain age or grade group is based on the arithmetic mean and the standard deviation. Most frequently called standard scores, they have advantages over such relative measures of placement as percentile scores and are thought by some²⁰ to be superior to the PC as derived scores of intelligence. The *Merrill-Palmer Scale of Mental Tests* uses standard score norms. Terman and Merrill present tables for the use of research workers and other persons in converting IQs obtained on the *New Revised Stanford-Binet Tests of Intelligence* into standard scores.²¹

III. DISTRIBUTION OF INTELLIGENCE

It is important that the teacher know something of the manner in which intelligence is distributed, if he is to make effective use of intelligence test results. The many reports of the distribution of intelligence show, however, that no single pattern of the distribution of intellectual ability can be expected to apply widely to different school situations. Typical groups of school children are not unselected, as might be supposed, but have been affected variously in their composition by many selective factors.

Intelligence can be conceived of both in terms of some such measure as the IQ and in terms of descriptions of the types of performance possible for persons of different intelligence levels. A distribution of intelligence quotients for an unselected group of children and the general descriptive terms used for different levels are presented here as an indication of the general distribution of intelligence.

Table X shows the distribution of intelligence quotients for a normal population. Figure 18 presents the same data graphically. It will be noted that sixty percent of the population fall within ten IQ points of the average IQ of 100.

²⁰ Francis N. Maxfield, "Trends in Intelligence Testing" *Educational Research Bulletin*, 15 134-41, May 13, 1936

²¹ Lewis M. Terman and Maud A. Merrill, *Measuring Intelligence*, p. 42. Houghton Mifflin Co., Boston, 1937.

On the average, one person in each 100 is in the very superior class or above and one person in each 100 is feeble-minded. About seven percent of the total may be considered as distinctly superior and seven percent as distinctly inferior. About one person in 400 is classified as a genius, while one in 400 is also found to be in the imbecile or idiot class. Persons at the highest level of feeble-mindedness, i. e., morons, are not uncommon in the lower grades of the school.

TABLE X
DISTRIBUTION OF INTELLIGENCE QUOTIENTS IN A NORMAL
POPULATION ²²

Classification	IQ	Percentages of all Persons
Near genius or genius	140 and above	0.25
Very superior	130-139	0.75
Superior	120-129	6.00
Above average	110-119	13.00
Normal or average	90-109	60.00
Below average	80-89	13.00
Dull or borderline	70-79	6.00
Feeble-minded Moron	50-69	0.75
Imbecile	25-49	0.19
Idiot	24 and below	0.06

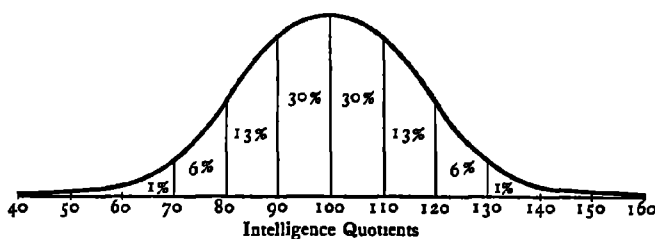


FIGURE 18 PERCENTAGE OF PERSONS IN A NORMAL POPULATION AT DIFFERENT
LEVELS OF INTELLIGENCE

²² Adapted from Peter Sandiford, *Foundations of Educational Psychology*, 1939, Longmans, Green and Co., New York, 1939.

IV. CLASSROOM USES OF GENERAL INTELLIGENCE TESTS

The need for information such as is provided by good intelligence tests is so great that no classroom teacher can be assured that his procedure with a given pupil is sound unless he has access to such exact information. The majority of group intelligence examinations must be used carefully and interpreted cautiously. Many of the group tests measure with sufficient accuracy for group comparisons, but thus far the results from even the best of them should not be taken too seriously when it comes to making individual comparisons. Some group tests permit the determination of mental ages and intelligence quotients. These derived scores, if based on group test results, are undoubtedly too unreliable for use in cases in which important decisions are involved, but are, in the main, sufficiently accurate for most school situations. If it is important that the mental age or the IQ of an individual pupil be obtained and only group tests can be used, not less than two forms of the test should be given at different periods. Even this procedure may result in conclusions which may be seriously erroneous. However, it need hardly be pointed out that a measure of intelligence based on a group test is better than anyone's guess on the subject. In any event, it is sound practice to test all very high and very low ranking pupils with a good individual intelligence test, for it is at the extremes of the distribution that intelligence test results are the least reliable and it is also at the extremes of ability levels that difficulty in effecting optimum adjustments of pupils is greatest.

For Individual Diagnosis. The intelligence test may prove especially valuable to the classroom teacher in assisting him to solve the problems relating to the unusual child. The pupil may be unusually bright, troublesome, dull, mischievous, or in some other way quite out of the ordinary. The teacher may wish to know whether this child's typical responses reveal his real general ability, and whether or not the judgments of his former teachers and supervisors are correct. Intelligence tests will give information not obtainable in any other way. The results may indicate that the child has ability hitherto unsuspected, or that his supposed

brightness is but a superficial pertness covering a really dull intellect. In a case of misconduct, they may show that the child's bad behavior comes from the failure of the teacher to keep busy a really brilliant intellect, thus allowing his superfluous mental activity to find its outlets in mischief and disorder.

The intelligence test, when given to an entire group, frequently uncovers a child of outstanding ability who has been content to go on with the group without revealing his real ability. Many such cases would never have been revealed, had it not been for the intelligence test. Such tests invariably uncover cases of overlapping in ability just as achievement tests reveal cases of overlapping in school achievement. They may reveal children in the fourth and fifth grades with the intelligence of normal seventh- or eighth-grade pupils, or fourth- and fifth-grade pupils who in intelligence are at the second- or third-grade level.

When children are discovered who are far in advance of their place in school, readjustments of work should be made to match their abilities. This may be accomplished by: (1) advancing them to a grade where their intelligence is given a real test, (2) placing them in rapidly moving classes, so that they may progress according to ability rather than by some fixed promotion scheme, or (3) declaring minimal requirements for the entire class in the units of work to be done and then expecting the brighter pupils to attack the various problems at higher levels and more intensively than could be expected of the class as a whole.

By proceeding along similar lines, the teacher will better understand the dull pupil because his difficulties can then be diagnosed more particularly and his strong points brought into relief. Quite frequently this results in an entire re-direction and reorganization of his instruction. In any case, the intelligence test will assist both in explaining difficult cases and in revealing unsuspected general strengths and weaknesses.

For Educational Guidance. The use of intelligence test results for educational guidance is similar to but goes far beyond their use for individual pupil diagnosis. Pupils can be much more effectively advised in their selection of courses

and of curricula, and, by the same token, courses can better be adapted to their needs, if information is available concerning their intellectual levels. Pupils may be better qualified for certain types of courses or curricula than for others, in terms of their levels of intelligence. Evidence now available from some tests concerning ability levels in several areas or types of performance makes even more significant than formerly the possibilities of using intelligence test results in this manner. The use of specific intelligence or aptitude tests in this connection is also of great value for educational guidance.

For Vocational Guidance. The dividing line between educational and vocational guidance cannot be clearly drawn, for the first merges gradually into the second. Whereas educational guidance is of primary concern in the elementary school, even there it has its vocational implications. Vocational aspects of guidance assume an increasingly prominent position as the pupil progresses through junior and senior high school and in many instances nears the end of his school career. Although intelligence and aptitude test results can be used with less confidence for vocational than for educational guidance, the information they furnish concerning the general and specialized intellectual abilities of pupils is of great value in vocational counseling.

For Class Analysis and Diagnosis. Viewed from the standpoint of the teacher, achievement tests and intelligence tests are supplementary devices. After the teacher has given achievement tests and compared his class with the norms in a given subject, he is still in danger of making false assumptions relative to the significance of these results unless he has available further information such as is furnished by intelligence tests. He may credit himself with an excellent job of teaching, when the innate brilliance of the group in his charge is such that, if they were given really adequate instruction, much superior results might have been achieved. Or it may be that the class falls considerably below the expected norm, and the teacher may feel that his teaching has proven a failure. This may also be an unwarranted assumption, for the class may be considerably below average in intelligence and cannot be expected to approximate the norms that

are set up for a class of average ability. It is evident then that there is a need for some means of determining approximately the intellectual ability of the class. Intelligence tests meet this need. By giving one or more such tests, the teacher can determine with a fair degree of accuracy whether his class is up to the normal expectation in ability to master school work.

The teacher is almost certain to encounter a greater number of technical difficulties in interpreting intelligence test results than he finds in using achievement tests. However, with practice and with the use of proper precautions he may come to have confidence that his conclusions are approximately correct for the class as a whole. He may, by comparing the results of several of these group mental tests, make a diagnosis of the ability of his class which will go a long way toward enabling him to determine what degree of proficiency he should reasonably expect in the teaching of subject matter. This knowledge will enable him to plan his work more intelligently and to lay plans for overcoming individual difficulties which might otherwise have been unsuspected.

V. CLASSROOM USES OF SPECIFIC INTELLIGENCE TESTS

Much of what has been said concerning the uses of general intelligence tests applies also to aptitude and readiness tests. However, the specific nature of these types of tests limits their significance to certain uses which are correctly made of general intelligence tests.

The aptitude test, measuring aptitudes for performance in such different fields as mechanics, music, art, and academic subjects, is valuable for individual pupil diagnosis, educational guidance, and vocational guidance. However, it has less significance for class analysis and diagnosis because it measures such specific abilities that individual pupil characteristics assume much greater importance than do characteristics of the class as a whole. The aptitude test is primarily suited to use for pupils of the junior high school or higher levels, for the general and non-specialized type of course in the elementary school is less well adapted to apti-

tude testing than are the more specialized courses of the high school and the college.

Readiness tests, available for reading and for arithmetic at the elementary school level, are useful for individual pupil diagnosis and for educational guidance but seem to have little significance for vocational guidance or class analysis and diagnosis. Such tests also have specific rather than general significance, so that the results from their use should be interpreted for the pupil as an individual rather than on the basis of the class group.

Aptitude or specific intelligence tests will be given attention in the consideration of various subject fields in later chapters of this volume. Their specific nature seems to make that treatment preferable to a more complete discussion of their uses at this point.

VI. CLASSROOM USES OF PERFORMANCE TESTS

Performance tests are less frequently a tool of the classroom teacher than of the educational or vocational counselor. Pupils who have visual, language, or physical handicaps which preclude reliable testing of their abilities by group intelligence tests should be tested by individual intelligence scales or performance tests. However, such instruments are somewhat less widely used by the classroom teacher than are group intelligence tests and aptitude and readiness tests, so they are given relatively little emphasis here. The uses of results from performance tests do not differ significantly from the uses of group intelligence test results except that performance tests furnish less accurate measures of general intelligence than do group and individual intelligence tests and therefore should be employed with caution.

VII. DERIVED MEASURES RELATING INTELLIGENCE AND ACHIEVEMENT

It has been pointed out in the preceding chapters that intelligence tests are designed to measure primarily innate or inherited abilities and that achievement tests are intended to measure the results of education and experience. In one

sense, then, intelligence tests can be considered as measuring *the ability to learn* or *the potentialities for achievement* and achievement tests can be considered as measuring *what has been learned*. It seems natural, therefore, that an effort should be made to discover how well the individual is living up to his potentialities by comparing his performances on intelligence and achievement tests.

Two general procedures have been used for this purpose—those based on quotients and those based on differences. Three specific methods are discussed here. Only the first, which is discussed most fully, has come into wide use, but the other procedures are briefly presented so that the student may better grasp the problems involved in a reliable comparison of ability with achievement. As is pointed out later, measures of this type are highly questionable in their use with individual pupils, and even for use with pupil groups they must be interpreted with care and with regard for the many variables which condition their use if important pupil adjustments are to be made on the basis of the results.

Accomplishment Quotient (AQ). The accomplishment quotient or achievement quotient, also sometimes called the accomplishment or achievement ratio, seems first to have been used in 1920 in the *Illinois Examination*, which was composed of parts measuring both intelligence and achievement. Franzen elaborated upon its use²³ during the same year.

The accomplishment quotient represents the relation between the educational level (EA) and mental maturity (MA) or between the relative educational development (EQ) and relative brightness (IQ) of a pupil. Therefore, the formula for the AQ is, in several adaptations,

$$AQ = 100 \frac{EA}{MA} = 100 \frac{EQ}{IQ} = 100 \frac{\frac{EA}{CA}}{\frac{MA}{CA}},$$

where EA, MA, and CA indicate respectively the educational, mental, and chronological ages of the pupil expressed

²³ Raymond Franzen, "The Accomplishment Quotient" *Teachers College Record*, 21 432-40, November 1920.

in months and EQ and IQ designate respectively his relative educational development and brightness.

For example, if a child has a mental age of ten years (120 months) and an educational age of nine years (108 months), his

$$AQ = 100 \frac{9-0}{10-0} = 100 \frac{108}{120} = 90.$$

If a pupil's achievement (EA) is in keeping with his ability to learn (MA), his AQ will be 100. Obviously, the indication of an AQ below 100 should be that the child is not working to capacity and an AQ of more than 100 should be impossible. However, a study of highly motivated instructional drives on certain subject matter²⁴ shows that an AQ of more than 100 is attainable. It is certain, however, that no one can achieve at more than 100 percent of his capacity, so it would appear that such accomplishment quotients result from norms on achievement tests which are spuriously low.

There is evidence to show that higher accomplishment quotients are more frequently obtained in particular grade groups by the intellectually inferior than by the intellectually superior pupils.²⁵ This probably is true largely because of the fact that the instructional levels of most schools are geared to the average and inferior pupils and that the curriculum frequently does not have enough "top" adequately to interest and motivate superior pupils. Therefore, an AQ below 100 may indicate poor effort, a high IQ, or both, and an AQ of more than 100 may indicate unusual effort, a low IQ, or both.

Another weakness of the AQ is its low reliability,²⁶ which results from the fact that a ratio between two measures which are themselves not highly reliable for this comparison (EA

²⁴ W E Lessenger, *Motivation and the Accomplishment Quotient Technique*. University of Iowa Studies in Education, Vol III, No 2 University of Iowa, Iowa City, 1925

²⁵ Harl R Douglass and C L Huffaker, "Correlation between Intelligence and Accomplishment Quotient" *Journal of Applied Psychology*, 13 76-80, February 1929.

²⁶ J Crosby Chapman, "The Unreliability of the Difference between Intelligence and Educational Ratings" *Journal of Educational Psychology*, 14 103-8, February 1923.

and MA, or EQ and IQ) cannot be highly reliable because the quotient of two unreliable measures is less reliable than either of the measures. In defense of these ages and quotients, it should be said that they have satisfactory degrees of accuracy for their normal uses but that they probably are not satisfactorily reliable in the not-wholly-defensible ratios used in obtaining the AQ.

A sound conclusion, growing out of the above and other more technical evaluations of the AQ, seems to be that its use with individual pupils is probably not justified but that it can satisfactorily be used for groups of pupils.

Educational and Mental Indices. Another method for measuring the difference between achievement and ability to achieve is that devised by Pintner and Marshall.²⁷ A mental index and an educational index, similarly based on ranges of 100 points and having 50 as averages, are compared. If the result from the application of the formula

$$\text{Difference} = \text{Educational Index} - \text{Mental Index}$$

is positive, the individual is making good use of his abilities; if the difference is negative, he is not working to capacity.

Index of Studiousness. An index of studiousness which attempts to relate ability to performance in the classroom has been proposed.²⁸ In its simplest form, this measure is the difference between a pupil's rank in his class on intelligence and on achievement, as they are measured by standardized tests. The index of studiousness is practically limited in comparable application to pupils within a class or instructional group and is recommended by its originator primarily for use in the high school.

TOPICS FOR DISCUSSION

1. Under what conditions, if any, do you think classroom teachers should be responsible for giving and scoring intelligence tests?
2. Propose a program to be followed in a school with respect to the recording and use of intelligence quotients or other derived scores of intelligence.

²⁷ Rudolf Pintner and Helen Marshall, "A Combined Mental-Educational Survey" *Journal of Educational Psychology*, 12 32-48, January 1921.

²⁸ Percival M. Symonds, *Measurement in Secondary Education*, pp. 521-25. The Macmillan Co., New York, 1928.

3. Discuss fully the most commonly used measures of mental maturity and brightness.
4. To approximately what age does mental growth continue?
5. Does a person's intelligence quotient remain constant throughout his life? Give evidence to support your answer.
6. What do you understand to be the significant differences between the intelligence quotient and the personal constant?
7. Discuss the use of percentile scores and standard scores for the interpretation of intelligence test results.
8. How is intelligence distributed among the population as a whole?
9. List and discuss some of the ways in which intelligence test results are useful in the classroom.
10. What does the accomplishment quotient attempt to show? Discuss its defects and proper uses.
11. If a child has a CA of 12-6 and an MA of 15-10, what is his IQ? If his EA is 14-6, what is his AQ?
12. If the parents of a sixth grade child who had a CA of 13-6 and an MA of 10-1 called upon you to discuss his poor work in arithmetic, what would you tell them about the causes of the child's deficiency in arithmetic?

SELECTED REFERENCES

- Bingham, Walter V., *Aptitudes and Aptitude Testing*. New York. Harper and Brothers, 1937.
- Boynton, Paul L., "Intelligence and Intelligence Tests" *Encyclopedia of Educational Research*, pp. 622-34. New York: The Macmillan Co., 1941.
- Broom, M. E., *Educational Measurements in the Elementary School*, Chapters XI-XII. New York McGraw-Hill Book Co., Inc., 1939.
- Cattell, Psyche, "The Heims Personal Constant as a Substitute for the IQ." *Journal of Educational Psychology*, 24 221-28, March 1933.
- Conrad, Herbert S., "Norms." *Encyclopedia of Educational Research*, pp 773-80. New York: The Macmillan Co., 1941.
- Cureton, E. E., "The Accomplishment Quotient Technic." *Journal of Experimental Education*, 5 315-26, March 1937.
- Dearborn, Walter F., *Intelligence Tests · Their Significance for School and Society*. Boston Houghton Mifflin Co., 1928.
- Dickson, Virgil E., *Mental Tests and the Classroom Teacher*. Yonkers-on-Hudson, N Y World Book Co., 1923.
- Freeman, Frank N., *Mental Tests: Their History, Principles, and Applications* (Revised Edition). Boston. Houghton Mifflin Co., 1939.
- Kelley, Truman L., *Scientific Method · Its Function in Research and Education*, Chapter III. New York: The Macmillan Co., 1932.
- Madsen, I. N., *Educational Measurement in the Elementary Grades*, Chapters V-VI. Yonkers-on-Hudson, N. Y. World Book Co., 1930.

- Odell, C. W., *Educational Measurements in High School*, Chapter XV. New York: The Century Co., 1930.
- Orleans, Jacob S., *Measurement in Education*, Chapter 3. New York: Thomas Nelson and Sons, 1937.
- Peterson, Joseph, *Early Conceptions and Tests of Intelligence*. Yonkers-on-Hudson, N. Y.: World Book Co., 1925.
- Pintner, Rudolf, *Intelligence Testing* (New Edition). New York: Henry Holt and Co., 1931.
- Richardson, M. W., "The Logic of Age Scales." *Educational and Psychological Measurement*, 1 25-34, January 1941.
- Sandiford, Peter, *Foundations of Educational Psychology*, Chapter V. New York: Longmans, Green and Co., 1939.
- Smith, Henry L., and Wright, Wendell W., *Tests and Measurements*, Chapter XIX. New York: Silver Burdett and Co., 1928.
- Stoddard, George D. (Chairman), *Intelligence Its Nature and Nurture*. Thirty-Ninth Yearbook of the National Society for the Study of Education, Parts I and II. Bloomington, Ill.: Public School Publishing Co., 1940.
- Symonds, Percival M., *Measurement in Secondary Education*, Chapter IV. New York: The Macmillan Co., 1928.
- Terman, Lewis M., *The Intelligence of School Children*. Boston: Houghton Mifflin Co., 1919.
- Webb, L. W., and Shotwell, Anna Markt, *Testing in the Elementary School*, Chapters V-VII. New York: Farrar and Rinehart, Inc., 1939.

CHAPTER XI

USING PERSONALITY INSTRUMENTS IN PUPIL GUIDANCE

The following aspects of personality and its measurement are discussed in this chapter.

- a.* Nature of personality.
- b.* Methods of personality measurement.
- c.* Nature and measurement of attitudes.
- d.* Nature and measurement of interests.
- e.* Significance and measurement of emotional adjustment.
- f.* Measurement of total personality.

Teachers are expected to understand their pupils, and through this understanding to increase the efficiency of their teaching. To the teacher of a few decades ago, all pupils were essentially alike. The modern teacher should have a knowledge of child psychology and the nature of individual differences in intelligence, achievement, and all other aspects of behavior. Many teachers give too little attention to the personality aspects of child behavior, preferring to work with the more readily observable and more tangible phases of behavior such as those treated in the chapters on achievement and intelligence testing. It is probable, also, that teacher-education institutions have too infrequently provided teachers in training with adequate instruction concerning pupil personality in a functional sense. Wherever the fault may lie, attention is increasingly being directed toward the more effective adjustment of the school to the needs of the child and of the child to life. Thus efficient teaching demands more than a chance and casual acquaintance with personality testing techniques.

I. THE NATURE OF PERSONALITY

Man has for centuries been aware of differences among individuals and has made many attempts to classify them. As early as 2000 B.C., Theophrastus divided men into thirty

universal types,¹ of which the dissimulator, the flatterer, the chatterer, and the rustic are representative. Hippocrates, several centuries B.C., distinguished persons of the sanguine, choleric, melancholic, and phlegmatic characteristics and explained these various types of temperaments by excesses of the bodily fluids or "humors" he called blood, yellow bile, black bile, and phlegm respectively.² Palmistry, phrenology, numerology, graphology, and others of what are frequently called "pseudo-sciences" have long made claims concerning personality which have largely been disproved by scientific experimentation.

More recently, Kretschmer divided men by their physical characteristics into four types distinguishable by certain general personality characteristics,³ and Berman emphasized the influence of secretions from the endocrine or ductless glands upon personality.⁴ Jung distinguishes extrovertive and introvertive types of individuals,⁵ and his classification has to a considerable degree found its way into popular usage.

More recently still, however, psychologists have increasingly turned their attention to the study of and attempts to measure personality. The concept of types evidenced in most of the early and many of the rather recent attempts to evaluate personality has largely been abandoned by modern personality testers, for personality types are inconsistent with the "normal curve" distribution which has been found to apply to personality traits as well as to intelligence and achievement.

Personality was at one time thought to be largely if not entirely the result of biological inheritance. However, most authorities today prefer the view that it is the resultant of both hereditary and environmental factors.⁶ Psychoanalysts

¹ Richard Aldington (Editor), *A Book of Characters from Theophrastus* E. P. Dutton and Co., New York, 1924

² Laurance Frederic Shaffer, *The Psychology of Adjustment*, p. 284. Houghton Mifflin Co., Boston, 1936

³ Ernst Kretschmer, *Physique and Character* Harcourt, Brace and Co., New York, 1925

⁴ Louis Berman, *The Glands Regulating Personality* The Macmillan Co., New York, 1921

⁵ C. G. Jung, *Psychological Types* Translated by H. G. Baynes Harcourt, Brace and Co., New York, 1923

⁶ Willard C. Olson, "Personality." *Encyclopedia of Educational Research*, p. 786. The Macmillan Co., New York, 1941.

believe that many of the personality difficulties found among adults are caused primarily by experiences, in many cases forgotten by the adults, during early childhood. If personality characteristics are the result in significant measure of the environment, which seems a justifiable conclusion, it is important for the teacher to be alert to the influence of the school in shaping the personality of the child as well as to its potentialities for correcting the maladjustments which pupils may have acquired prior to school entrance.

Definitions of Personality. "Personality" is the most inclusive term which can be used in the discussion of human behavior. Psychologists are not in complete agreement concerning the meaning of the term, but they recognize that personality describes more fundamental types of human behavior than the surface evidences by which the man on the street evaluates it. In general, psychological definitions of personality explain what personality is in terms of the types of human behavior thought to contribute to it. Psychologists agree roughly upon these components of personality, but they usually resort to indirect methods of defining the term.

Shaffer states that the "personality traits of an individual are his persistent habits toward making certain types of adjustments rather than other kinds."⁷ Traxler considers the term to include the "sum total of an individual's behavior in social situations."⁸

These statements concerning personality seem to describe as well as possible in a non-technical manner what personality is. They perhaps represent the most meaningful view of personality for teachers and other persons who are not technical workers in the field of personality study. It should be kept clearly in mind that the behavior of the individual is controlled by his personality and at the same time furnishes the evidence by which his personality can in part be evaluated.

Aspects of Personality. If personality is most satisfac-

⁷ Shaffer, *Op cit* p 132

⁸ Arthur E Traxler, *The Use of Tests and Rating Devices in the Appraisal of Personality*. Educational Records Bulletin No. 23, p. 4. Educational Records Bureau, New York, March 1938

torily described at present in terms of how it is constituted, it is understandable that approaches to personality study and measurement have been largely in terms of personality traits. Psychologists divide personality into many areas for study. However, the aspects of personality useful to the teacher can well be listed under fewer headings, although any classification must be largely arbitrary. The phases of personality treated in this chapter are grouped under the headings of attitudes, interests, emotional adjustment, and total personality. It is believed that these are the areas of greatest present significance to the teacher.

Although the preceding statements include intellectual and physical traits as components of personality, these traits are not generally considered when personality measurement is undertaken. They are measured by different techniques in established areas of testing, so, although the psychology of personality rightly deals with their findings and no one should lose sight of the contributions of intellectual and physical traits to an individual's development, these areas are not of direct concern here.

II. TECHNIQUES OF PERSONALITY MEASUREMENT

Personality is measured by several different types of approaches. Among those most commonly used are (1) free association, (2) direct observation of behavior, (3) rating scales, and (4) personal reports. Although all of these methods can be used by an intelligent classroom teacher, it is probable that observation of behavior and personal reports are the methods most practicable and useful in the typical classroom. Each of these methods will be discussed briefly in this section of the chapter. In the later sections, various methods of measurement will be discussed in terms of their uses for the measurement of attitudes, interests, emotional adjustment, and total personality. However, as the personal report method predominates, the other methods will be dealt with almost entirely at this point.

Personality testing is probably the newest area of measurement which bears directly upon the work of the classroom teacher. Although achievement and, to a less extent, in-

telligence, are subject to quantitative measurement and now have rather widely accepted technical terminologies which aid in the interpretation of testing results, such is not the case for personality testing. Many of the results from personality tests must be interpreted in qualitative rather than quantitative terms, and the area has practically no derived scores such as the educational age, the intelligence quotient, etc., which have commonly accepted meanings. The effect of this situation is that personality test results must be interpreted largely in terms of the special types of derived scores and norms provided for the particular test used and then frequently by qualitative rather than quantitative statements.

Association Methods. An association method was one of the earliest to be employed in the measurement of behavior, for it apparently was first used by Galton as early as 1879.⁹ Its development has occurred mainly since 1910 in the modern sense, however.

Two association methods are now being quite widely used in the study of personality: (1) the free association method and (2) the projective method. The first approach has long been used, but the projective method is quite a recent development.

Free associations are established when the person to whom a word is spoken responds with the first word which enters his mind. Another free association procedure is based on ink blots, to which the subject is to respond by telling what he is reminded of or what he sees in each. Both the nature of the responses and the manner in which they are given to situations of these types furnish considerable evidence to the experienced psychologist on which to base inferences concerning emotional disturbances in the subject.

Projective methods make use of an observational procedure, but they are considered here because it is the manner in which the individual responds to a given situation which makes his behavior meaningful. The child is presented with some such materials as sand, clay, toys, or pictures, and his use of the material presented is carefully observed by the psychologist. Much is revealed to the experienced observer

⁹ Francis Galton, "Psychometric Experiments" *Brain*, 2 149-62; July 1879.

concerning the conscious and even unconscious motives, attitudes, interests, and needs of the individual by this approach.

Man has doubtless for centuries drawn inferences concerning the behavior of others from their speech and actions. The associative methods are not, then, new. However, the attempt to apply systematic procedures is relatively new, and the projective method in particular is now under critical examination by many students of personality.

Observation of Pupil Behavior. Several different methods based on the observation of pupil behavior have been suggested and successfully applied. They all probably require an ability in the use of such methods which few teachers have but which most can acquire. Teachers are perhaps prone to observe group behavior rather than individual pupil behavior, except in very unusual situations, whereas it is the individual who becomes the center of reference in these observational techniques. Untrained teachers also make use of their own interpretations of events they observe, whereas objectivity is attained only by a rather rigid account of what actually occurred. These and other characteristics of observational methods highly useful in the study of pupil personality and adjustment make it inadvisable for inexperienced teachers to attempt to make more than experimental use of them until some experience in observation has been acquired.

Greater insight concerning individual pupil behavior and a greater comprehension of child behavior in general should result in any teacher who is alert to observe and report significance in behavior situations. Consequently, he should better be able to recognize and understand the personality problems of his pupils and to aid the pupils in attaining more satisfactory adjustment.

The two most common observation procedures in the school are (1) directed observation and (2) the anecdotal method. The first, because the observation is directed toward a particular pupil or pupil group under specified conditions, is a laboratory rather than a classroom procedure. The second, however, uses the results from observations of pupil behavior made at any time, and therefore is definitely a classroom method of evaluation.

Teachers have doubtless for generations used the anecdotal method in their spare-time discussions about pupils. However, its first use as an evaluative instrument was probably as recent as 1928.¹⁰ The anecdotal record is an objective description by the teacher of a significant occurrence or episode in the life of the pupil. Unless a situation has sufficient meaning to a teacher who is alert to the underlying motives governing human behavior to bring it definitely to his attention, it probably is not of sufficient significance for inclusion in the anecdotal record.

An anecdotal record must be carefully, although not laboriously, prepared if it is to be of value. The anecdote is a highly objective brief of what occurred in a situation in which a pupil behaved in a sufficiently unusual manner to make the incident meaningful. It may consist of an objective narrative of the incident only or it may consist of the narrative, an impartial interpretation of the occurrence, and even a recommendation for guidance of the pupil concerned. If interpretations and recommendations are given, however, they should be distinguished from the original description so that their nature is clearly apparent to a person reading the anecdotal record. The anecdotal record has great value only when it is made cumulative by the addition of new anecdotes as meaningful situations arise and are observed and recorded by the teacher or some other school officer.

Rating Scales. Rating scales are widely used in the evaluation of pupil personality. In this procedure, the teacher or some other person intimately acquainted with the pupils rates them on personality traits in terms of the manner in which the individuals have impressed the rater. Obviously the raters should know quite intimately the pupils they are rating. Most rating methods suffer in accuracy because of the fact that some raters tend to be too lenient and others too critical. They are less accurate for use with intangible traits, upon which different raters usually vary rather widely, than in the more readily observable characteristics such as neatness, cleanliness, etc.

Widely used among rating techniques are the graphic rat-

¹⁰ D. A. Robertson, (Chairman), "Report of Subcommittee on Personality Measurement" *Educational Record*, 9 53-68, Supplement No. 8, July 1928.

ing scale and variations of that form. In this procedure, the rater places a check mark at a certain position on a line to indicate his evaluation of the person he is rating. The line may be divided into five (or some other number of) sections designated superior, good, average, poor, and inferior, or meaning may be given to positions on the scale by other and more definitely descriptive terms. Again, there may be designations at occasional intervals beneath the line to indicate specifically for each trait varying evidences of its possession by the person being rated. Several personality rating scales are discussed later in this chapter.

Personal Reports. The personal report method makes use of what are variously called scales, inventories, questionnaires, blanks, etc. The responses are given or the instruments are filled out by the pupils themselves. As many of the items on these instruments request highly personal responses, the personal report method of measuring personality suffers from the fact that pupils sometimes reply as they think they should reply rather than in terms of their true reactions to the various items. Most persons are hesitant in revealing their inner personalities to other persons freely. In fact, the customs of civilized society place something of a premium upon the ability to hide or disguise emotions, likes and dislikes, attitudes, and other reactions in many situations. Therefore, it is not surprising that pupils sometimes fail to answer personality test items truthfully. Despite this major weakness, personal report instruments for the measurement of personality are of considerable value in their classroom uses.

III. MEASUREMENT OF ATTITUDES

A significant portion of the teacher's time in the classroom is directly or indirectly devoted to the development in pupils of desirable social attitudes and modes of behavior. Illustrations are the emphasis in the school upon good citizenship, cooperation with others, intellectual honesty, and the scientific attitude. Furthermore, many courses in the school attempt to develop attitudes which are in many cases more specific than those mentioned above. For example, the

teacher of English attempts to develop favorable attitudes toward correct usage and good literature in his pupils, and the teacher of civics to develop democratic ideals and belief in democratic institutions. Lists of course objectives for the elementary school include many such attitudes, ideals, or beliefs which the school strives to develop or improve in the child. It seems important, then, that teachers be conversant with instruments for measuring attitudes of various types.

The Nature of Attitudes. Thurstone defines an attitude as "the sum total of a man's inclinations and feelings, prejudice or bias, preconceived notions, ideas, fears, threats and convictions about any specific topic."¹¹ An attitude is a state of readiness which exerts a directive, and sometimes a compulsive, influence upon an individual's behavior.

Attitudes may be either general or specific. For example, a person who has a general attitude of liberalism may behave in a highly conservative manner in a particular situation in which his personal welfare may be threatened. An attitude of conservatism is general, but an attitude toward a certain person is specific. This brief indication of the nature of attitudes will furnish the student sufficient background concerning the psychological characteristics of attitudes for the brief consideration of measuring instruments presented in this chapter.

Methods of Attitudes Measurement. Attitudes are measured by several different methods, among the most common of which are the interview and the scale or questionnaire. As the teacher ordinarily has more use for the attitudes scale than for the interview, only brief mention will be made of the interview as a device for determining attitudes and opinions.

The Interview. The method of determining attitudes by the use of the interview is similar to the method of general interviewing discussed in the following chapter. However, the interview for purposes of attitudes measurement is usually restricted to rather direct and somewhat standardized questioning on the particular issue toward which attitudes are being measured.

¹¹ L. L. Thurstone and E. J. Chave, *The Measurement of Attitude*, pp. 6-7. University of Chicago Press, Chicago, 1929

Attitudes Scales. The two series of attitudes scales most widely used and known are the *Thurstone Scales for the Measurement of Social Attitudes* and the *Generalized Attitudes Scales* devised by Remmers and his associates. However, they will not be discussed here because they are intended for use with high school and college students. The *Pressey Interest-Attitude Test*, illustrated on page 33, is an instrument which in effect measures both attitudes and interests.

Sample items from the *Around the World* attitudes inventory for pupils from the sixth to the tenth grade are presented in an accompanying illustration. These items show a type of approach to attitudes measurement suitable for intermediate grade and junior high school pupils. This inventory measures attitudes toward various phases of international relations, war, patriotism, and agencies for peace.

SAMPLE ITEMS FROM "AROUND THE WORLD" ATTITUDES INVENTORY¹²

I. YES OR NO?

Below are a number of statements, some of which are true and some are false. Underline YES if you think a statement is true. If you think it is not true, underline NO.

- | | |
|------------------------------------------------------------------------------------------------|--------|
| 1. Most foreigners are less intelligent than Americans | YES NO |
| 7. In most American homes there are things made out of material that came from other countries | YES NO |
| 10. In modern warfare people who live far from the fighting are often in great danger | YES NO |

Another attitudes scale for the elementary school level makes use of a considerably different approach. This is the *Personal Attitudes Test for Younger Boys*, intended for use with boys from nine to eighteen years of age in the discovery of behavior and emotional problems, to throw light on certain traits of personality, and to evaluate programs of character and religious education. The accompanying illustration shows a few sample items which are presented only in part because of space limitations. In the inventory itself, the "How I Feel" columns are reproduced in exactly the same manner under the two additional headings "How Most

¹² Adelaide T. Case and Paul M. Limbert, *Around the World*. Published by Association Press, 1932.

EXCERPT FROM A PERSONAL ATTITUDES TEST FOR YOUNGER BOYS¹³

	HOW I FEEL				
	Dis- like	Rather Not	Don't Care	Like Some	Like a Lot
25 Doing physical exercises					
26 Always being brave.....					
27 Learning to drive a car.....					

Boys Feel," and "How I Think I Ought to Feel." The pupil is to indicate his answers by encircling the response under each of the three headings which represents his attitude.

IV. MEASUREMENT OF INTERESTS

Attention of classroom teachers has increasingly turned of late years to pupil interests, as a result of the emphasis now placed upon the adaptation of the school offerings to the abilities, needs, and interests of pupils. Furthermore, the vocational and avocational interests of children have increasingly received attention during the past couple of decades as a means of aiding the pupil in the selection of his school courses and curricula and his life vocation. The school obviously cannot adapt its offerings to pupil interests and guide pupils in their selection of courses in terms of interests if the nature of those interests is unknown.

The Nature of Interests. Interests today are most often classified in terms of the objects and activities from which the individual obtains satisfaction.¹⁴ Thus, a person is interested in football, but cares very little for tennis, or he is interested in music but is not interested in the drama. It is in this non-technical manner of considering interests that measurement in this field can be most meaningful for the teacher.

¹³ *A Personal Attitudes Test for Younger Boys*. Published by Association Press, 1928.

¹⁴ Douglas Fryer, *The Measurement of Interests*, p. 15. Henry Holt and Co., New York, 1931.

Methods of Interests Measurement. Interests are subject to measurement both by informal instruments and by standardized interests inventories. Brief mention will be made here of informal testing methods. Interests inventories will be treated somewhat more fully.

Informal Measurement of Interests. Educational literature of recent years includes many reports of interests studies in a variety of school subjects and areas of behavior. Among the fields for which such studies have appeared in considerable number are reading interests in books, magazines, and newspapers; play interests; interests in the movies and radio; and interests in various subjects of the elementary school and high school. Reference to such sources will furnish the teacher much information concerning interests of various pupil groups.

However, the teacher can obtain direct information concerning the interests of his pupils by informal methods. Questioning individual pupils and class groups about their interests is a simple procedure, and one productive of considerable information. The teacher may, however, have the pupils write about their interests or list them without discussion. Again, he may prepare and distribute to the pupils a list of books, of magazines, of recreational activities, or of any one of a number of other types of objects and activities and then ask the pupils to check those in which they are interested. In any of these procedures, it is wise to limit the investigation of interests to one area rather than to attempt a complete inventory of pupil interests at one time.

Interests Inventories. It is perhaps because of the fact that inventories of pupil interests in objects and activities are so easy to make that standardized inventories usually deal with interests from the standpoint of their predictive or diagnostic values for important types of behavior. The result is that certain types of instruments which are perhaps best classified as interests inventories do not differ greatly from attitudes scales and that other types are very similar to, or in effect may be, adjustment inventories. The Pressey *Interest-Attitude Test*, a brief excerpt from which appears on page 33, is an illustration of the first type.

The *Dunlap Academic Preference Blank* makes use of expressed interests of junior high school pupils in a relatively small number of persons, concepts, objects, and activities to furnish data useful for pupil guidance and classification and for determining relative scholastic aptitudes.¹⁵ Scores

EXCERPT FROM DUNLAP ACADEMIC PREFERENCE BLANK¹⁶

The purpose of this paper is to learn what your interests are in various things, people, places, and activities in which students are usually interested. Be perfectly honest in marking the paper. The results will have no effect on your grades. The necessary directions are given below. Read these directions carefully. You will have plenty of time, but work rapidly.

Mark every item.

After each word or phrase you will see four answer spaces, marked L I D U

L is for Like

I is for Indifferent or Do not care

D is for Dislike or Do not like

U is for Unknown

If the word or phrase names something you like to read about, to have, or to work with, fill in the space under L. If you are indifferent or do not care, fill in the space under I. If you don't like what the word suggests, fill in the space under D. If you do not know what the word means, fill in the space under U.

For example, one boy marked the next three items as follows

Castor oil	L	I	D	U
	1			
Taking a bath	L	I	D	U
	1			
Candy	L	I	D	U
	1			

Now you mark these items to show how you feel about them.

Castor oil	L	I	D	U
	1			
Taking a bath	L	I	D	U
	1			
Candy	L	I	D	U
	1			

Do not omit any of the items.

Be careful to mark each item only once. You may re-read the directions at any time. Do not turn over this paper until you are told to begin.

	L	I	D	U		L	I	D	U		L	I	D	U	
The Sphinx	1	I	I	I		The solar system	31	I	I	I	Hemp	31	I	I	I
Historic tales	2	I	I	I		Drama	32	I	I	I	Greyfriars Bobby	32	I	I	I
Sanitation	3	I	I	I		Tuberculosis	33	I	I	I	Bandaging wounds	33	I	I	I
Leaf Ericson	4	I	I	I		Indian territories	34	I	I	I	Civil Service jobs	34	I	I	I
Wigwags	5	I	I	I		Picnics	35	I	I	I	Ice Patrol of the Arctic	35	I	I	I
Honolulu	6	I	I	I		The corn belt	36	I	I	I	Oil wells	36	I	I	I
Legends	7	I	I	I		Fearful	37	I	I	I	Merrimack and Monitor	37	I	I	I
Expletives	8	I	I	I		Hyphens	38	I	I	I	Analysis of sentences	38	I	I	I
Pioneers	9	I	I	I		Alexander Graham Bell	39	I	I	I	Nathan Hale	39	I	I	I
City libraries	10	I	I	I		Ping pong	40	I	I	I	Band concerts	40	I	I	I
Pagodas	11	I	I	I		Sidney, Australia	41	I	I	I	Pygmies	41	I	I	I
Ben Hur	12	I	I	I		Pied Piper	42	I	I	I	Aesop's Fables	42	I	I	I
Multiplicands	13	I	I	I		Product in multiplying	43	I	I	I	Division problems	43	I	I	I
The Alamo	14	I	I	I		Lexington and Concord	44	I	I	I	Surrender of Yorktown	44	I	I	I
Jugglers	15	I	I	I		Bulldogs	45	I	I	I	Alligators	45	I	I	I

¹⁵ Jack W. Dunlap, *Manual of Directions Dunlap Academic Preference Blank*, p. 6. World Book Co., Yonkers-on-Hudson, N. Y., 1940.

¹⁶ Jack W. Dunlap, *Dunlap Academic Preference Blank*. Published by World Book Co., 1940.

resulting from scoring the test in different combinations of item weights are predictive of academic success in seven aspects of junior high school instruction and general achievement, and the blank also furnishes mental ability and intellectual alertness scores. The illustration on page 256 reproduces the directions and some of the items of the blank.

The Brainard and the Stewart-Brainard *Specific Interest Inventories* analyze tendencies essential to various vocations by means of items classified into twenty groupings based on

EXCERPT FROM BRAINARD SPECIFIC INTEREST INVENTORY,
BOY'S FORM¹⁷

Note hour and minute at which this was begun -- -- -- --

Put a circle around one number after each question, as in practice.

Ph.

How do you like—

		Dislike		N		Like
1	To dig in the ground, make a garden, plant trees, etc. r	1	2	3	4	5
2	To move furniture, beds, boxes, or other heavy objects? ..	1	2	3	4	5
3	To cut the lawn, rake the yard, clip grass, shrubs, etc. ?	1	2	3	4	5
4	To use your muscle on things hard to unscrew, things that stick, things that have to be pounded or pried loose?	1	2	3	4	5
5	To climb on ladders, trees, or poles, to balance or hold on with one hand? ..	1	2	3	4	5

Look at the items you have just marked. Did you circle one number after each question? If you liked one part of a question, such as "make a garden," better than the other parts, did you underline those words? If you disliked one part more than your number shows, did you put an X through it? Use a 3 if you are not sure what to put down, but be sure to put a circle around some number after each question.

the general nature of the interests. Four separate forms are provided—for boys and for girls from 10 to 16 years old and for men and for women over 16 years of age. Combinations in different ways of the scores made on the various parts result in scores indicative of interests in twenty different types of occupations. The accompanying illustration of the items from the boys' form which deal with physical activities indicates the general nature of the inventory.

V. MEASUREMENT OF EMOTIONAL ADJUSTMENT

Every individual faces the problem of adjusting himself to a none-too-benign environment. Persons who are successful in adapting themselves to their environments are

¹⁷ Paul P. Brainard, *Specific Interest Inventory, Form B*. Published by The Psychological Corporation, 1932.

well adjusted; those who fail in this adaptation become maladjusted. The school seeks to improve the adjustment of its pupils by furnishing them important learning opportunities and experiences. However, it must go beyond learning in the subject matter sense and attempt to bring about the best possible form of adjustment between the individual and his environment in terms of his total personality.

The measurement of adjustment is an extremely comprehensive task. In its broad sense, such measurement implies the use of all types of devices which will furnish information concerning the child and his backgrounds of heredity and environment. The discussion of adjustment in this section applies primarily to emotional adjustment. Although this is a fundamentally important issue, because of the fact that maladjustment seems to have consequences of great importance in the emotional life of the individual, the measurement of emotional maladjustment should not be regarded as the sole approach to this problem. The discussion of Chapter XII deals with adjustment in a somewhat broader sense than the treatment given in this section.

Causes and Symptoms of Maladjustment. Maladjustment may arise when an individual is frustrated in the satisfaction of his fundamentally important aims, motives, or goals. It is the result of a lack of balance between the difficulties the individual encounters in his environment and his ability to meet the difficulties successfully. The underlying causes may be of many types, and frequently they are very elusive. Frustration itself is a result, not a cause. The effects, or results, are much more readily determined than are the causes. Symptoms of maladjustment may fairly readily be observed by the teacher who has insight into pupil behavior, but the determination of causes underlying maladjustment is often a task for the clinical psychologist. Although some alleviation of maladjustment may be accomplished without knowledge of its causes, effective remediation depends upon a knowledge of and ability to cope successfully with the true causal factors.

Methods of Adjustment Measurement. The importance of an awareness by the teacher of existent emotional malad-

justments in his pupils should be apparent from the preceding discussion. Such recognition of maladjustments should be accompanied by evidence concerning their nature, and, if possible, their causes. Adjustment inventories serve the first two purposes of pointing out the existence of and nature of existing maladjustments quite adequately in many instances, but they probably do not accomplish the third purpose, of discovering the causes of maladjustments. They frequently, however, furnish evidence which will greatly facilitate further study of maladjusted pupils in the attempt to determine causes and then to eliminate them.

Two general procedures are probably most often used in the measurement of adjustment—rating scales and personal report blanks. Each of these methods is discussed briefly and illustrated by a few representative instruments in the following pages.

Rating Scales. Two rating scales which are of major use in locating maladjusted pupils are briefly commented upon and illustrated here. Although these scales have the same general purpose as the personal report blanks discussed in the next section of this chapter, the two types of adjustment measures differ greatly in method.

The *Haggerty-Olson-Wickman Behavior Rating Schedules* are illustrated by the few following items. Although this rating scale is similar in general appearance to the graphic

EXCERPT FROM HAGGERTY-OLSON-WICKMAN BEHAVIOR
RATING SCHEDULES¹⁸

					Score
25. Is he even-tempered or moody?	Stolid, Rare changes of mood (3)	Generally very even tempered (1)	Is happy or depressed as conditions warrant (2)	Strong, and frequent changes of mood (4)	Has periods of extreme elations or depressions (5)
26. Is he easily discouraged or is he persistent?	Melts before slight obstacles or objections (3)	Gives up before adequate trial (3)	Gives everything a fair trial (1)	Persists until convinced of mistake (2)	Never gives in, Obstinate (4)
27. Is he generally depressed or cheerful?	Dejected Melancholic, In the dumps (3)	Generally dispirited (4)	Usually in good humor (1)	Cheerful Animated, Chirping (2)	Hilarious (5)

¹⁸ M E Haggerty, W C Olson, and E K Wickman, *Haggerty-Olson-Wickman Behavior Rating Schedules*. Published by World Book Co., 1930.

rating scale, it differs in that the two extremes do not necessarily represent the most and the least desirable situations. Instead, the numbers 1 to 5, variously spaced for different items, indicate in descending order the relative desirability of the stated condition.

EXCERPT FROM HAYES SCALE FOR EVALUATING SCHOOL BEHAVIOR ¹⁰

Directions for Using this Scale

Following is a list of habits which children 10 to 15 years old have been found to show. No one child could have all the habits listed, but is certain to have a considerable number of them.

Draw a circle around the T, F or U before each item to indicate. (T) you believe the statement is true of the child being rated, (F) you believe the statement is not true of the child being rated, (U) you are uncertain whether the statement is true or not true of the child being rated. Be sure to draw a circle about one letter and one only for every item in the list. Two samples are given below:

- (T) F U usually accepts responsibility when the occasion arises
T (F) U often wastes time

Circle the following items in a similar manner

I

- | | | | | |
|---|---|---|-----|-------------------------------------------------------------|
| T | F | U | 1. | often does little things to make others happy |
| T | F | U | 2 | usually thinks of consequences both to self and others |
| T | F | U | 3 | usually accepts responsibility when the occasion arises |
| T | F | U | 4 | often shares with others |
| T | F | U | 5. | usually does his share in any group activity |
| T | F | U | 6 | often "plays hookey" from school |
| T | F | U | 7. | usually does the work expected of him |
| T | F | U | 8 | usually defends his friends only when they are in the right |
| T | F | U | 9. | usually makes friends easily |
| T | F | U | 10. | often starts fights |
| T | F | U | 11. | usually quickly forgives wrongs done to him |
| T | F | U | 12 | often uses vulgar or profane words |
| T | F | U | 13 | usually eats lunch with a group |

The *Hayes Scale for Evaluating the School Behavior of Pupils Ten to Fifteen* is another rating scale used either by the teacher in rating his pupils or by the pupils in obtaining self-ratings. Directions for using the scale and a few sample items are given in the accompanying illustration. The

¹⁰ Margaret Hayes, *A Scale for Evaluating the School Behavior of Children Ten to Fifteen*. Published by The Psychological Corporation, 1933.

results of a simple scoring procedure can be used in preparing a school behavior profile, in locating maladjusted pupils, and in spotting the behavior areas in which maladjustment seems to exist.

Another type of instrument which is in effect a rating scale is the Baker "*Telling What I Do*" tests. The accompanying illustration from the advanced level test for pupils in Grades 7 to 9 illustrates the method of measuring pupil behavior. Scores can be obtained for the following areas of behavior: (1) school, (2) home, (3) play, (4) social, and (5) ethical-moral.

EXCERPT FROM BAKER "TELLING WHAT I DO" TEST²⁰

On this sheet you will find many things about yourself. Some of these things are known about you already, but we want you to tell us yourself.

Each exercise has three answers. You are to draw a line under the one answer to each exercise that most nearly tells what you do. Put the letter of the answer in the parenthesis at the end of the line.

Underline only one answer to each exercise. Take the one that most nearly fits you. Be honest with yourself. Underline what you really do, even if it is not what you know you should do.

There are eighty exercises. Answer all of them. Take your time, and think over each exercise carefully. It should take you at least half an hour, or longer, to do all the exercises as you really should.

-
- | | | |
|--------------------------|--------------------------|------------------------------|
| 1. Tardy for school | | |
| a. Never tardy | b. Often tardy | c. Tardy once in a while () |
| 2. When I lose a game | | |
| a. I just quit | b. Don't care if I lose | c. Try harder next time () |
| 3. Eating | | |
| a. Usually hurry | b. Eat very fast | c. Eat slowly () |
| 4. When I meet strangers | | |
| a. Like to meet them | b. Don't care about them | c. They bore me () |
| 5. If I borrow | | |
| a. I never pay back | b. Pay back right away | c. Pay when asked () |
-

Personal Report Blanks. By far the majority of adjustment inventories make use of the personal report method, by which pupils are asked to give answers to a variety of questions. The considerable quantity of adjustment inventories and the wide variety of response methods they use precludes any more comprehensive treatment here than brief descriptions and illustrations of a few of them.

An illustration from the *Bell Adjustment Inventory*, an instrument for measuring (1) home, (2) health, (3) social, and (4) emotional adjustment, was given on page 33, so this inventory will not be discussed further here.

²⁰ Harry J. Baker, "*Telling What I Do*," Advanced Form. Published by Public School Publishing Co., 1930.

Sample items from several parts of the girls' form of the Rogers *Test of Personality Adjustment* are given herewith to illustrate procedures used in measuring adjustment of elementary school children. This inventory, for use with girls from nine to thirteen years old, is devised to measure adjustment of the girl toward other children, toward her family, and toward herself. The comparable form for boys is not illustrated here.

EXCERPTS FROM ROGERS TEST OF PERSONALITY ADJUSTMENT
FOR GIRLS²¹

NUMBER ONE

Suppose that just by wishing you could change yourself into any sort of person. Which of these people would you wish to be? Write a "1" in front of your first choice, a "2" in front of your second choice, and a "3" in front of your third choice

- | | |
|----------------------------|----------------------|
| (a) _____ a housewife | (n) _____ a fireman |
| (b) _____ a teacher | (o) _____ a poet |
| (h) _____ a business woman | (t) _____ an actress |
| (l) _____ an aviator | (y) _____ a salesman |
| (m) _____ a captain | (z) _____ an artist |

Is there any other sort of person you would like to be? If there is, write it here _____

NUMBER FOUR

Read the sentences below, and the questions that follow them. If the answer to a question is "yes," put a check mark (✓) on "yes." If the answer is "no," put a mark on "no." If the true answer is somewhere in between yes and no, put the mark where it will be most true.

1. Mary is the prettiest girl in school.

Am I just like her?

Yes								No
-----	--	--	--	--	--	--	--	----

Do I wish to be just like her?

Yes								No
-----	--	--	--	--	--	--	--	----

NUMBER FIVE

In the questions that follow, put a mark (✓) in front of the line that is the true answer, unless it tells you to do otherwise.

1. How well can you play ball?

- (a) _____ can't play ball at all.
 (b) _____ can play a little bit.
 (c) _____ can play pretty well
 (d) _____ best player in my class.

²¹ Carl R. Rogers, *A Test of Personality Adjustment for Girls*. Published by Association Press, 1931.

The *Aspects of Personality* inventory measures the temperament and personality traits of children in Grades 4 to 9 by the use of items of the type shown in the accompanying illustration. The inventory yields scores which can be translated into percentiles on an ascendance-submission, an extroversion-introversion, and an emotionality scale.

EXCERPT FROM ASPECTS OF PERSONALITY ²²

SECTION III

III

- | | | |
|------------------------------------------------------|----------------------------|----------------------------|
| 1. I like to go to the movies. | <input type="checkbox"/> S | <input type="checkbox"/> D |
| 2. I think most children like to make fun of me..... | <input type="checkbox"/> S | <input type="checkbox"/> D |
| 3. I get angry about nothing. | <input type="checkbox"/> S | <input type="checkbox"/> D |
| 4. I get so angry I can't talk. | <input type="checkbox"/> S | <input type="checkbox"/> D |

The *Guess Who Test*, illustrated by the sample given below, is intended for use in measuring a child's reputation among his fellows. The test, for use from Grade 5 to Grade 8, requests pupils to list their classmates who particularly fit the brief portraits presented to them. It is possible to obtain a total reputation score for each pupil in a class from the results.

EXCERPT FROM GUESS WHO TEST ²³

Here are some little word-pictures of children you may know. Read each statement carefully and see if you can guess who it is about. It might be about yourself. There may be more than one picture for the same person. Several boys and girls may fit one picture. Read each statement. Think over your classmates and write after each statement the names of any boys or girls who may fit it. If the picture does not seem to fit anyone in your class, put down no names but go on to the next statement. Work carefully and use your judgment.

1. Here is the class athlete. He (or she) can play baseball, basketball, tennis, can swim as well as any, and is a good sport.

²² Rudolf Pintner, et al, *Aspects of Personality* Published by World Book Co., 1937

²³ *Guess Who Test* Published by Association Press, 1930.

VI. MEASUREMENT OF TOTAL PERSONALITY

No very definite line of distinction can be drawn between the instrument to be discussed here and the types of adjustment inventories presented in the preceding section. However, the instrument presented here perhaps measures total personality rather than various aspects of personality. In its provision of means for determining a personality quotient (PQ), essentially comparable in significance to the IQ and EQ, there is at least an implication that it measures personality more broadly than do most of the currently available adjustment inventories.

McCall has developed an *Inter-Trait Rating Scale* which yields personality quotients in each of 43 areas of behavior which reflect personality and also an average PQ. The scale can be used for persons 12 years of age and older as a self-rating instrument or for obtaining ratings by friends. The essential feature of McCall's procedure is to compare the amount of each trait possessed by an individual with the amount of some objectively measurable trait, such as intelligence, he possesses.

The first two rows of the accompanying copy of the *Inter-Trait Rating Scale* are filled in with ratings for an hypothetical individual who has an IQ of 115, as a basis for showing how the scale is used. The rater feels that the individual is lower in accuracy than in intelligence, so he places a "—" in the second column. He estimates that his judgment on the individual's accuracy is about 40 per cent of certainty, so he writes "40" in the third column. The PQ column is then filled out by taking half of the percentage of certainty and, because the sign in the second column is negative, subtracting it from the IQ of 115 to obtain a PQ of 95. On adaptability, the plus rating with a 30 percent degree of certainty indicates that 15 should be added to the IQ of 115, to net a PQ of 130. The average of the 43 separate PQs then becomes the general PQ for the individual.

McCall believes that the embarrassment which sometimes arises when one person is asked to rate a friend in the friend's presence will not arise with this rating scale. He says, in

PERSONALITY INSTRUMENTS IN PUPIL GUIDANCE 265

McCALL INTER-TRAIT RATING SCALE ²⁴

Traits	Above or Below Intelligence	Percent of Certainty	Personality Quotients (½ the % plus IQ)
Accuracy	—	40	95
Adaptability	+	30	130
Appearance			
Cheerfulness			
Conscientiousness			
Cooperativeness			
Courage			
Courtesy			
Decisiveness			
Democracy			
Effectiveness			
Enthusiasm			
Foresight			
Generosity			
Happiness			
Healthiness			
Independence			
Industriousness			
Initiative			
Leadership			
Likeableness			
Loyalty			
Open-Mindedness			
Orderliness			
Originality			
Persistence			
Pleasing Voice			
Poise			
Progressiveness			
Punctuality			
Refinement			
Reliability			
Self-Confidence			
Self-Control			
Sense of Humor			
Sincerity			
Sociability			
Sympathy			
Tact			
Thoroughness			
Tolerance			
Truthfulness			
Vivacity			

Average

²⁴ William A. McCall, *Measurement*, p. 315. The Macmillan Co., New York, 1939.

commenting upon the manner in which some friends rated him.²⁵

Since they could not rate him down in accuracy without rating him up in intelligence or up in adaptability without rating him down in intelligence there was no particular embarrassment to them or him in these ratings, although the author does not see himself as others see him at certain points. They were not asked to state whether the author was very dull or very intelligent or very accurate or inaccurate, nor even to state how much difference there is between his intelligence and his accuracy.

TOPICS FOR DISCUSSION

1. In what way is a knowledge of personality measurement procedures valuable to the teacher?
2. What is meant by personality? How do psychologists and laymen differ in their conceptions of personality?
3. Briefly characterize two association methods of evaluating behavior
4. Indicate the nature of observation procedures for the evaluation of personality
5. What is the nature of graphic rating scales?
6. How are personal reports used in personality measurement?
7. Briefly indicate the nature of attitudes. Of what concern are they to the teacher?
8. Indicate the nature of one or two attitudes scales for use in the elementary school.
9. What is the nature of interests? How are pupil interests of significance to the teacher?
10. Discuss the two major procedures which are used in the measurement of interests
11. What are the causes and symptoms of emotional maladjustment? Which are easier to recognize? Why?
12. Indicate some of the methods by which pupil adjustment is measured.
13. Discuss how total personality is measured by one technique. What is the PQ?

SELECTED REFERENCES

- Allport, G. W., *Personality A Psychological Interpretation*. New York: Henry Holt and Co., 1937
- Appel, Kenneth E., and Strecker, Edward A., *Practical Examination of Personality and Behavior Disorders*. New York: The Macmillan Co., 1936.
- Bruceknor, Leo J., and Melby, Ernest O., *Diagnostic and Remedial Teaching*, Chapter XIII. Boston: Houghton Mifflin Co., 1931.

²⁵ Ibid p. 314.

- Freeman, Frank N, *Mental Tests Their History, Principles, and Applications* (Revised Edition), Chapter VIII. Boston Houghton Mifflin Co, 1939
- Fryer, Douglas, *The Measurement of Interests*. New York · Henry Holt and Co, 1931.
- Groves, Ernest R., *Personality and Social Adjustment*. New York : Longmans, Green and Co, 1931
- Hartshorne, Hugh, and May, Mark A, *Studies in Deceit*. New York : The Macmillan Co., 1928
- Hartshorne, Hugh, May, Mark, and Shuttleworth, F K, *Studies in the Organization of Character* New York The Macmillan Co., 1930.
- Hartshorne, Hugh, May, Mark, and Maller, Julius B, *Studies in Service and Self-Control*. New York The Macmillan Co, 1929
- Hunt, Thelma, *Measurement in Psychology*, Part VI. New York Prentice-Hall, Inc, 1936.
- Lee, J Murray, *A Guide to Measurement in Secondary Schools*, Chapter IV. New York D. Appleton-Century Co, Inc, 1936.
- Olson, Willard C, "Personality" *Encyclopedia of Educational Research*, pp. 785-95. New York The Macmillan Co, 1941.
- Sandiford, Peter, *Foundations of Educational Psychology*, Chapter VI. New York Longmans, Green and Co., 1939.
- Shaffer, Laurance Frederic, *The Psychology of Adjustment*. Boston : Houghton Mifflin Co, 1936
- Sherman, Mandel, *Mental Conflicts and Personality*. New York Longmans, Green and Co, 1938
- Sheviakov, G. V., and Friedberg, Jean, "Use of Interest Inventories for Personality Study." *Journal of Educational Research*, 33 692-97; May 1940.
- Stagner, Ross, "Attitudes" *Encyclopedia of Educational Research*, pp 69-75 New York The Macmillan Co., 1941
- Stogdill, Emily L, and Herndon, Audell, *Objective Personality Study · A Workbook in Applied Mental Hygiene* New York Longmans, Green and Co., 1939.
- Symonds, Percival M., *Diagnosing Personality and Conduct*. New York . D Appleton-Century Co, Inc, 1931
- Thorpe, Louis P, *Personality and Life*. New York : Longmans, Green and Co, 1941.
- Thorpe, Louis P., *Psychological Foundations of Personality*. New York · McGraw-Hill Book Co, Inc, 1938
- Traxler, Arthur E, *The Nature and Use of Anecdotal Records* (Revised). Educational Records Supplementary Bulletin D. New York : Educational Records Bureau, January 1939.
- Traxler, Arthur E, *The Use of Tests and Rating Devices in the Appraisal of Personality* Educational Records Bulletin No 23. New York : Educational Records Bureau, March 1938
- Wallin, J. E. W., *Personality Maladjustments and Mental Hygiene*. New York McGraw-Hill Book Co, Inc., 1935.

- Watson, Goodwin, "The Specific Techniques of Investigation Testing Intelligence, Aptitudes, and Personality." *The Scientific Movement in Education* Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XXX, pp. 357-73. Bloomington, Ill.: Public School Publishing Co., 1938.
- Weedon, Vivian, "A Technique for Determining Interest." *Educational Research Bulletin*, 13 191-97, 231-34; November 14 and December 12, 1934.
- Wrightstone, J. Wayne, *Appraisal of Newer Elementary School Practices*, Chapter XII New York Bureau of Publications, Teachers College, Columbia University, 1938.
- Zachry, Caroline B., *Personality Adjustments of School Children*. New York Charles Scribner's Sons, 1929.

CHAPTER XII

THE USE OF OTHER TECHNIQUES AND TOOLS IN PUPIL GUIDANCE

This chapter considers rather broadly the following types of techniques and tools useful in bringing about desirable adjustment of pupils to life:

- a.* Adjustment and guidance.
- b.* Cumulative pupil records as guidance tools.
- c.* Individual pupil profiles in guidance
- d.* Class analysis charts in guidance
- e.* Other techniques in the adjustment of pupils.

The underlying purpose of all educational measurement is to obtain information useful in effecting the better adjustment of the pupil to school and out-of-school life. This chapter supplements the preceding chapter by indicating briefly some of the tools and techniques which furnish information in composite form from various of the possible sources. The chief purpose of this chapter is to illustrate a few of the methods, for the possibilities in this type of approach to pupil guidance are too great for adequate treatment in less than an entire volume. However, some of the values to be realized by this broad approach to pupil evaluation may be brought home to the teacher by this brief treatment.

I. EDUCATION AS ADJUSTMENT

Adjustment and Guidance. Education is concerned with the problems of aiding the individual in making the best possible adjustments to his environment, present and future. Thus, practically, education *is* guidance and adjustment. Unfortunately, many institutions and teachers are so concerned with the teaching of subject matter that they forget their responsibilities for the development of personality. Furthermore, many of the adjustments required of pupils in their education and personality development are so complicated that teachers frequently find themselves ill-equipped

to direct them. The teacher, or guide, presumably is one who has been over the route before, and thus knows something of the experiences and the difficulties to be encountered. However, personality and character traits are so many-sided that it is too much to expect an individual teacher to be able to give the pupil equally good guidance in making adjustments in all phases of his environment. Even the most richly informed teacher must have supplementary information if this guidance is to be even reasonably effective. It is at this point that certain educational measuring devices make many helpful contributions to the teacher and to the individual pupil himself.

School Responsibility for Pupil Adjustment. The school through its administrative officers and its teachers faces a very real responsibility for the proper guidance and adjustment of its pupils. Correct guidance, if generally practiced, would pay big dividends in a more effective society and in happier and better adjusted individuals. Not everyone requires assistance in making the required adjustments to life, but those who do usually need it badly. All agencies engaged in rendering social service face a definite responsibility for the development of a sane and comprehensive guidance program. Common types of guidance are vocational, educational, mental, physical, civic, moral, avocational or cultural, and social. While these forms of guidance are frequently administered by different agencies, they are highly related and involve many identical problems. For the school and the teacher these problems are particularly vital.

II. GUIDANCE IN ADJUSTMENT

Guidance of the Whole Child. The general objective of guidance is to aid pupils in more wisely purposing, planning, executing, and evaluating all of the varied activities and interests which engage their attention. The aim of all guidance is to assist individuals in becoming happy and efficient human beings. Guidance is more than leading, conducting, regulating or directing persons, things, or activities. These terms imply too much external and too little internal responsibility for the progress of the individual's own education.

Accordingly, guidance must provide personal help that is offered in such a manner as to create in each individual the desire and the ability to carry on by his own efforts. The accomplishment of these objectives of guidance is not an easy task.

Guidance is an extremely inclusive term, and one which requires that the whole child be taken into account. Jones has the following to say concerning guidance :

Gradually, but surely, we have come to realize that guidance is not something that concerns only a part of the individual, nor does it deal merely with a part of his life. The need is for "whole" child guidance.¹

Importance of Individual Differences. The recognition that no two persons are alike is basic to effective guidance. Each case constitutes a complex adjustment problem, whether or not the person is actually maladjusted, which must be treated as an entity. Individuals differ greatly in their characteristics. They differ in their abilities to learn and in their special abilities and interests. They differ in the amount and type of their learnings. They differ in their physical makeups and abilities. They differ in their ethical and moral standards and their attitudes. They differ, it may be said in summary, in the thousands of characteristics they possess as results of their differing biological inheritance and their differing environments.

The Adjusted Pupil. The properly adjusted individual is one who has learned to fit himself smoothly and without warping his personality into the social group of which he would normally form a part. Proper adjustment of the individual implies that the different human qualities are properly balanced and combined in right relations into an integrated whole—the total personality. This total personality involves the happy integration of the physical, intellectual, moral, social, aesthetic, and religious natures of each individual. This point of view treats normality as the criterion of adjustment. This in no sense means that the teacher is to disregard the importance of individual differ-

¹ Arthur J. Jones, *Principles of Guidance* (Second Edition), p. 427. McGraw-Hill Book Co., Inc., New York, 1934.

ences or deviations from the norm. As a matter of fact, it places a greater premium on their recognition.

From the standpoint of the school, the adjusted pupil is the individual who is making normal and regular educational progress as compared with other individuals of similar capacities and traits. He is associated with groups of comparable mental level. He is adapted morally and socially to the individuals he meets in his daily life. He is physiologically adjusted in his sight, hearing, speech, and general health. The purpose of defining adjustment in this way is to make it apparent that the important thing is the process of fitting the individual into surroundings which are reasonably normal for him, rather than merely finding his absolute level of abilities or deficiencies. The mentally handicapped twelve-year-old child would be badly adjusted if placed with normal twelve-year-olds. The child with a speech defect may be maladjusted if placed with children with normal speech, but may readily adjust himself so far as his personality is concerned if he is placed with individuals by whom he is not submerged. This is a concept of prime importance to the teacher. The use of all available techniques and instruments for the collection and interpretation of guidance information which will make possible this adjustment is an obligation which each teacher must assume.

Basic Approaches to Pupil Adjustment. As the basis for effecting the changes in the educational program necessary to bring about better pupil adjustment in the school, many types of accurate and detailed information concerning the individual are needed. In this connection a clear distinction should be made between facts and opinions. Although subjective estimates of a pupil's achievement during his earlier educational history may be of significant value in a guidance program, they are far less useful than are objective scores from valid and reliable measuring instruments, other things being equal. Prominent among the basic approaches to pupil adjustment, therefore, is the matter of the availability of sufficiently accurate and extensive information about the individual. Of almost equally great importance is the need for a sympathetic and tolerant interpretation of this guidance information.

Among the more important sources of information for guidance purposes is the complete record of the individual's educational history, including the studies pursued, achievement in these subjects, participation in extra-curricular activities, acquisition of special honors and awards, and regularity of school attendance. Racial characteristics, the home life, and the social and economic conditions in which the pupil lives are also worthy of investigation. Data on health and physical conditions are frequently available from school records. Scores on intelligence tests, specific aptitude tests, and ratings on vocational interest and personality blanks can usually be obtained.

III. CUMULATIVE PUPIL RECORDS AS ADJUSTMENT TOOLS

An adequate system of cumulative pupil records is almost essential if a program of pupil guidance is to be effective. Many schools apparently keep no cumulative pupil records other than of background facts concerning the pupil and his parents and of scholastic success. Such records are wholly inadequate for anything more than a highly mechanical system of dealing with pupils. On the other hand, some schools have comprehensive and even elaborate systems of cumulative records which provide for the recording of a wide variety of data concerning each pupil in a cumulative record folder. Many types of variations between these extremes are also found.

No attempt is made here to catalog all of the types of information cumulative records typically contain. It is sufficient to indicate that the records should contain information about family background and environment, personal history, health, personality, intelligence, special abilities, school progress, scholarship, achievement test performances, extra-curricular activities, employment, educational plans, and vocational ambitions. Some record systems provide for the recording of data on all or most of these points on a record card or folder, and also for the filing of certain types of other data, such as anecdotal records and case studies, in the cumulative record folder.

35. 37 YEAR		1925		1926		1927		1928		1929		1930		1931	
NAME		1925		1926		1927		1928		1929		1930		1931	
36. AGE	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
38. SEX	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
39. RACE	White	Black	Other	White	Black	Other	White	Black	Other	White	Black	Other	White	Black	Other
40. BIRTHPLACE	White	Black	Other	White	Black	Other	White	Black	Other	White	Black	Other	White	Black	Other
41. EDUCATIONAL PLANS	High School	College	University	High School	College	University	High School	College	University	High School	College	University	High School	College	University
42. EDUCATIONAL RECORD	High School	College	University	High School	College	University	High School	College	University	High School	College	University	High School	College	University
43. NUMBER OF SIBLINGS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
44. NUMBER OF SIBLINGS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
45. NUMBER OF SIBLINGS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
46. NUMBER OF SIBLINGS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
47. NUMBER OF SIBLINGS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
48. PHYSICAL DISABILITIES	None	Stomach	Stomach	Stomach	Stomach	Stomach	Stomach	Stomach	Stomach	Stomach	Stomach	Stomach	Stomach	Stomach	Stomach
49. HEALTH RECORD	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
50. HEALTH RECORD	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
51. INTERESTS REPORTED	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps	Stamps
52. SOCIAL ADJUSTMENTS	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
53. SOCIAL ADJUSTMENTS	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
54. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
55. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
56. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
57. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
58. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
59. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
60. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
61. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
62. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
63. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
64. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good
65. PERSONALITY	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good	Good

FIGURE 19 SAMPLE OF AMERICAN COUNCIL ON EDUCATION CUMULATIVE RECORD FORM FOR ELEMENTARY SCHOOL PUPILS 2

It is impossible here to discuss at all adequately the values and uses of the cumulative record in pupil guidance and adjustment. However, it should be apparent that the mere availability of such a variety of information for all pupils in a school as is shown in the accompanying sample cumulative record would be of great value. Such records are

² Margaret W. Moore, *Cumulative Record Form for Elementary Schools*, pp. 6-7. American Council on Education

useful to administrators, to guidance workers, and to teachers as a basis for careful analyses in cases of maladjustment or disciplinary difficulties, and on others of the many occasions requiring or at least making desirable comprehensive information about particular pupils.

IV. THE USE OF TEST RESULTS IN THE ADJUSTMENT OF PUPILS

The adjustment tools with which this book is most directly concerned are tests and other evaluation techniques subject to use by the teacher in the guidance of his pupils. The types of information concerning a pupil which are shown in the above section to be important if not essential to adequate pupil guidance can in some cases be obtained only by the full-time guidance worker with the assistance of various specialists. However, the analytical uses of test results illustrated below are easily possible for the classroom teacher and the results can be of significant aid to him and benefit to the pupils.

Profile Charts. Profile charts are provided with many standardized tests of general achievement and many diagnostic tests in order to show differences in achievement levels graphically. The charts, frequently providing places for various total and part scores and their graphical representation, often appear on the front covers of the test booklets or on pupil answer sheets. It is sometimes recommended by test authors that the cover of the booklet, which carries such information as the pupil's name, the name and form of the test used, and the date on which it was given, in addition to scores and the profile, be torn off the used booklet and filed in the pupil's cumulative record folder or elsewhere for future reference and use.

The two following illustrations of profiles are representative of the types of charts provided with most general achievement tests. The first illustration shows the method of using a profile for results from a single test and the second shows use of the profile for indicating pupil progress over a period of several years.

A pupil profile chart which provides for the listing of

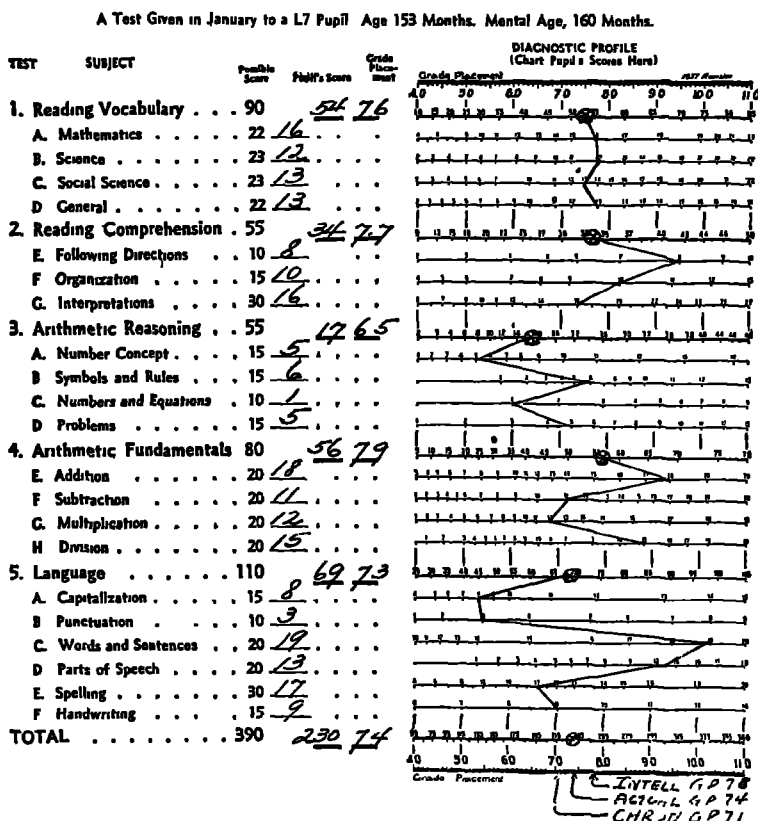


FIGURE 20 SAMPLE OF DIAGNOSTIC PROFILE CHART FOR PROGRESSIVE ACHIEVEMENT TESTS³

scores on an achievement test battery and for a profile showing achievement levels on the total test, on its five major areas, and on its twenty-one parts is that of the *Progressive Achievement Tests*. Relative strengths and weaknesses appear graphically for ready observation by the teacher.

The pupil whose record appears on the profile is in the lower half of Grade 7. He is a few months below the average age for the class chronologically and is several

³ *Manual of Directions Progressive Achievement Tests—Intermediate Battery*
p. 7. California Test Bureau, Los Angeles

months above the average in mental age. His achievement in both phases of reading and language is close to the class average. His slight acceleration in arithmetic fundamentals is about what might be expected of a child somewhat above the class average mentally. In arithmetic reasoning, however, the pupil is retarded about a year. Fluctuations of the profile for the various part scores in each major area show uniformity in reading vocabulary development, but tremendous differences appear for the various phases of language ability tested. Facts of these types are of value particularly in pointing out individual deficiencies which need special attention.

This brief interpretation of score data for a pupil illustrates the type of analysis which is possible, although only major differences are noted above. Evidence of this type is primarily analytic or diagnostic in the broad sense, but is not diagnostic in the specific sense of pointing out detailed types of scholastic deficiencies.

Evidence of pupil progress as measured by achievement tests over a period of years can be presented graphically by the use of a pupil profile chart. An illustration of this procedure is shown for a pupil tested for three successive years by the *Metropolitan Achievement Tests*. The pupil's average achievement is shown by the broken horizontal lines and his school grade level by the heavy horizontal lines. His mental and chronological ages at the times he was tested are shown under "Age," the years and months of mental age being indicated by circles.

The trends of each profile indicate the pupil's relative strengths and weaknesses at a given time. Outstanding characteristics of the profile based on testing when he was eleven years old, for example, are his above-average intelli-

gence ($IQ = 100 \frac{12-3}{11-0} = 100 \frac{147}{132} = 111$) and his below-

average age for the fifth grade, his relative superiorities in arithmetic fundamentals, geography, and spelling, and his relative weaknesses in reading vocabulary and arithmetic problem solving. Such evidence can well be used by the

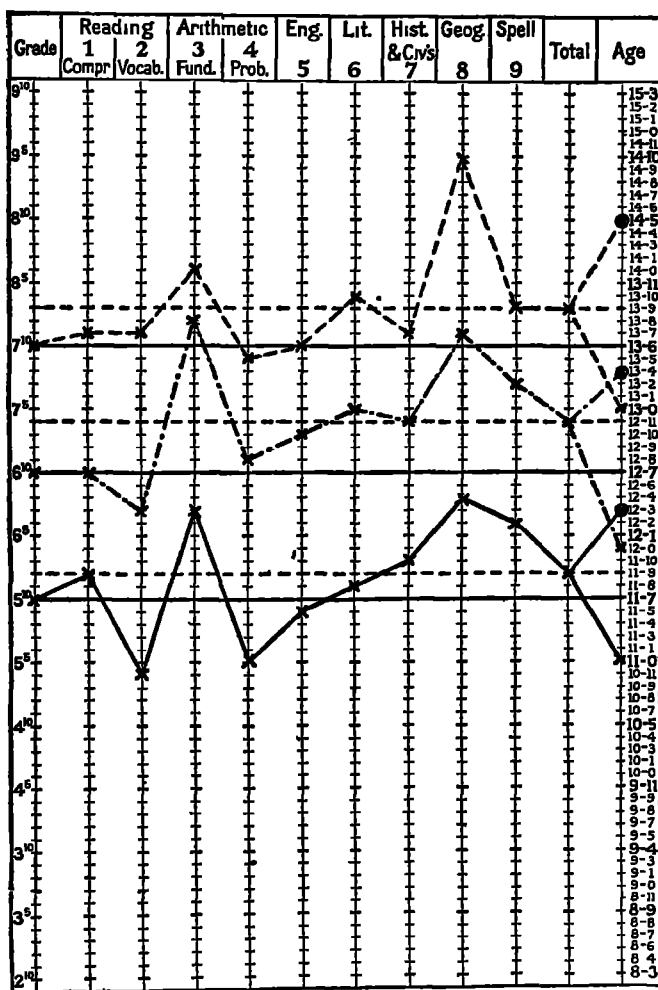


FIGURE 21 SAMPLE GRAPHIC RECORD OF PUPIL PROGRESS AS MEASURED BY METROPOLITAN ACHIEVEMENT TESTS ⁴

teacher to indicate major instructional areas to which special attention should be directed for the pupil.

For the three-year period, the rough similarities of the three profiles show fairly regular development in all aspects

⁴ Richard D. Allen, et al, *Supervisor's Manual Metropolitan Achievement Tests*, p. 32. World Book Co., Yonkers-on-Hudson, N. Y., 1935

of achievement. The pupil seems to have brought his reading and arithmetic abilities into better balance by significant progress in reading vocabulary and arithmetic problem solving, to have progressed at less than his average rate in spelling and history, and to have extended even further his previous superiority in geography.

Meaning attaches to such profiles in the manner suggested by these interpretations. When evidences of pupil progress are considered in relation to the other evidence the school should have about the pupil, additional meanings, and consequently added support for certain types of instructional and advisory emphases, should emerge.

One other test profile is illustrated here because it is fairly typical of profiles provided with some personality tests. For use with the *California Test of Personality*, it provides percentile positions along a scale for total adjustment, for self and social adjustment, and for twelve more specific areas of adjustment. Total adjustment at the 25th percentile and self and social adjustment respectively at the 20th and 45th percentiles show that the girl whose scores are represented on the chart is rather poorly adjusted, particularly in terms of self. Something of the pattern of her lack of adjustment can be inferred by a more complete analysis of the results than is given above.

Class Analysis Charts. Class analysis charts are valuable tools in the summarization of results from testing. Although such charts as are provided with standardized tests vary greatly, they usually provide a means of showing median achievement for the class or pupil group and the position of each pupil in the group in relation to age norms, grade norms, or both, for elementary school tests. High school tests more frequently provide for the graphical representation of median group performance and individual pupil status in relation to grade norms or percentile norms. The following illustration and discussion are based on a class analysis chart which is rather typical of those usually provided with general achievement test batteries.

The class analysis chart reproduced on page 282 gives an analysis of the results from the use of the *Metropolitan Achievement Tests* with a class of 40 pupils just completing

Name Mary Brown Grade H6 ..
 School Fairview St. .. Age 12 .. Last Birthday Sept 6 ..
 Teacher Miss Jones Date Nov 10, 1938 Sex: Boy-Girl

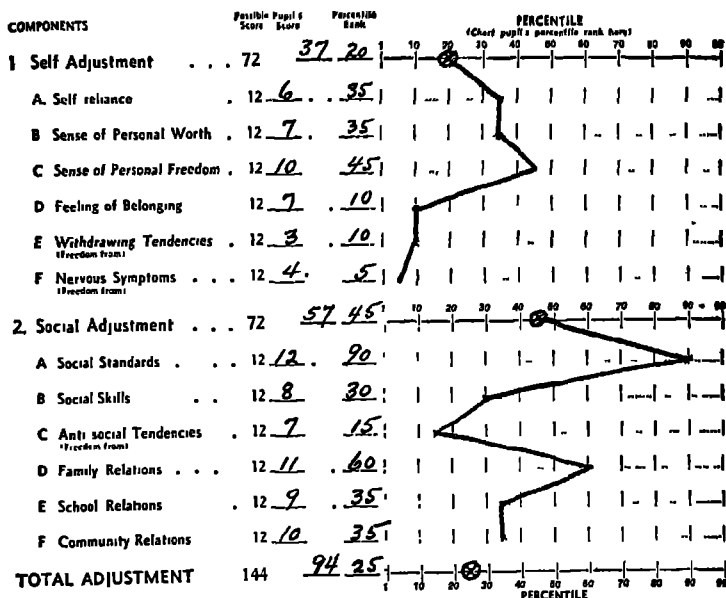


FIGURE 22. SAMPLE PROFILE CHART FOR CALIFORNIA TEST OF PERSONALITY⁵

the sixth grade. Median class standing is shown by the encircled crosses for achievement on the entire test and in the various subjects for which the test provides, as well as for chronological and mental ages. The line connecting the encircled crosses provides a profile of median achievement. Individual pupils are designated by identifying numbers placed at positions on the chart representing their scores. The distribution of intelligence quotients in columns 1 and 2 is related to but not really a part of the chart proper. At the bottom of the chart are shown the median achievement

⁵ Louis P. Thorpe, Willis W. Clark, and Ernest W. Tiegs, *Manual of Directions California Test of Personality—Elementary Series*, p. 7. California Test Bureau, Los Angeles.

in terms of grade equivalents and the months by which it is above or below the actual grade placement of the class at the time the test was given.

Data of these types are subject to many uses by teachers and supervisors. In this case, for example, the fact that the class medians are low for reading abilities, arithmetic problem solving, and spelling, but high for arithmetic fundamentals and geography, might indicate the desirability of giving more instructional emphasis the following year to those subjects in which relative deficiencies exist. Also, there is clear-cut evidence concerning the pupils who are outstandingly superior or inferior in each subject, and who might need special attention.

The range of ability by subjects is also useful. For example, pupil 36 at the 4⁴ and pupil 13 at the 8⁰ grade levels in arithmetic fundamentals are separated by four and a half years of ability in that area, yet pupil 36 has an IQ of 105 and pupil 13 an IQ of 97. The teacher should ask why pupil 36 shows such a tremendous disability, and perhaps be able to ascertain its cause and then to furnish adequate remediation. In contrast with arithmetic fundamentals, history shows a spread of ability of only two years—from 5⁰ for pupil 40 to 7⁰ for pupil 5. Teaching history should be easier than teaching arithmetic to this group, because of the greater homogeneity of the group in its history achievement, but the question might be raised as to whether or not progress of the superior pupils was not being sacrificed for the benefit of the weak pupils in the instructional procedures used or in the content of the course.

Only some of the more obvious types of interpretations of data on a class analysis chart have been mentioned, but these few perhaps illustrate the uses to which such an evaluative tool can advantageously be put.

Guidance Tests and Inventories. A set of instruments useful in various phases of guidance individually and in improving adjustment as a battery is the *Kefauver-Hand Guidance Tests and Inventories* for the junior high school level. The series is composed of an inventory of student plans and six separate guidance tests in the areas of: (1) educational, (2) health, (3) recreational, (4) social-civic, (5) vocational,

and (6) student-judgment. As the breadth of coverage implies, these instruments make use of various procedures for the measurement of knowledges, attitudes, interests, and adjustment. The accompanying illustration from the *Social-Civic Guidance Test* illustrates one of the techniques used.

EXCERPT FROM KEFAUVER-HAND SOCIAL-CIVIC GUIDANCE TEST⁷

PART III

Directions This is a test to see if you know how many of our citizens (boys, girls, men, and women) would have their well-being favorably affected (directly or indirectly) by each of the types of social action listed below

If you think less than 25% (less than one quarter) of our citizens would be favorably affected (directly or indirectly) by the social action, write 1 in the parentheses, as shown in the sample below

If you think between 25% and 75% would be favorably affected, write 2.

If you think more than 75% (more than three quarters), write 3

If you have no idea how many would be affected, write ? in the parentheses

Less than 25%	(1)
Between 25% and 75%	(2)
More than 75%	(3)
Don't know	(?)

Sample Providing regulations for the safety of deep-sea diving (1)

-
- | | |
|-----------------------------------------------------------------------------------------------------------------------------------------|-------|
| 1 Providing regulations which will permit the showing of only those pictures whose effects are beneficial to a majority of our citizens | () 1 |
| 2 Providing the broadcasting of only those radio programs whose effects are beneficial to a majority of our citizens | () 2 |
| 3 Providing regulations which will prevent the publication of misleading and untruthful advertisements | () 3 |
| 4 Providing better care for all our insane, crippled, and feeble-minded persons | () 4 |
| 5 Guaranteeing the constitutional rights of American citizens to all racial and political groups | () 5 |

V. THE USE OF OTHER TECHNIQUES IN THE ADJUSTMENT OF PUPILS

Certain evaluative tools closely similar to those which are discussed here were presented in the section of the preceding chapter which dealt with emotional adjustment. The tools presented there dealt primarily with personality phases of adjustment, however, whereas the interview and the case

⁷ Grayson N Kefauver, Harold C. Hand, and Virginia Lee Block, *Social-Civic Guidance Test*. Published by World Book Co., 1937.

study place practically no limits upon their scope in the use of those tools which is discussed here. Although personality involves behavior of the whole child, the adjustment inventories and methods treated in Chapter XI deal only indirectly with intelligence and achievement levels. The instruments briefly discussed below are concerned with the whole child in every aspect of what he is, what he does, and what he can do.

The Interview. The interview deserves only brief attention here, for the teacher is not directly concerned with it in its formal sense. The interview may, however, be informal and it may deal only with the areas of the child's interests, needs, background, etc., about which the teacher needs information. Even in such informal uses of the interview as may be of concern to the teacher, it is essential for best results that rapport be established between the teacher and the child. A frightened or an antagonistic pupil is not a good subject for an interview. Therefore, the teacher should give the same type of attention to the establishment of rapport which is necessary prior to the administration of individual intelligence tests. Pupils should not be questioned on many types of issues in the presence of a third party, for their responses might be less frank and spontaneous than if they were questioned in privacy.

In this broad sense, the interview is widely useful and flexible. However, it extends possibilities to the teacher for learning more about his pupils and consequently aids him in attempting to effect the best adjustment possible for each pupil.

Case Studies. The case study is a broad and comprehensive approach to the problems of pupil behavior. It should include extensive information about the present status of the pupil as well as about his past experiences and his family background. In fact, the case study may well draw upon many or even all of the types of information contained in adequate cumulative pupil records.

Usually there is a specific reason for making a case study. Such an approach may be used to gain a better understanding of a failing pupil, a disciplinary case, or a pupil who is poorly adjusted in one or another of many possible ways.

Cumulative pupil records should be consulted for pertinent information. The interview, discussed briefly above, may be used to supplement what information the cumulative record yields. Other channels may be used for the collection of other important data.

When information about the pupil is reasonably complete, the pertinent items should be organized into a meaningful whole and, in the formal case study, written up. Then should follow a tentative diagnosis of the causes of the undesirable behavior or performance of the child and the formulation of a plan of treatment. The final steps are remediation and a follow-up to determine whether it has been effective.

TOPICS FOR DISCUSSION

1. Do you accept the idea that in its broader aspects education is adjustment? Give your reasons
2. How are guidance and adjustment related? What is guidance in adjustment?
3. What is the responsibility of the school for the adjustment of its pupils?
4. Characterize the well-adjusted child. In what way is the "whole child" concerned?
5. Of what importance are individual differences of pupils in a guidance program?
6. Describe and explain a typical pupil profile chart
7. List and evaluate the importance of major sources of data for use in pupil guidance
8. What are some of the guidance uses of pupil profile charts?
9. Describe a typical class analysis chart and discuss its uses by the teacher
10. Briefly discuss the interview and the case study as methods of gathering and integrating information concerning pupils

SELECTED REFERENCES

- Allen, Richard D., *Inor Group-Guidance Series*, Volumes I-IV. New York Inor Publishing Co., 1934.
- Baker, Harry J., and Traphagen, Virginia, *The Diagnosis and Treatment of Behavior Problem Children*. New York The Macmillan Co., 1935
- Brown, Marian, and Brown, Martin V., "Records as Tools in the Study of Adolescents" *Educational Research Bulletin*, 15 207-15, November 11, 1936.

1. Jiederick, Paul B, "Evaluation Records." *Educational Method*, 15: 432-40, May 1936.
- Griffin, J. B, Laycock, S R, and Line, W., *Mental Hygiene*. New York American Book Co., 1940.
- Groves, Ernest R., *Personality and Social Adjustment*. New York Longmans, Green and Co, 1931.
- Hunt, Thelma, *Measurement in Psychology*, Chapter II. New York: Prentice-Hall, Inc, 1936
- Jones, Arthur J, *Principles of Guidance* (Second Edition). New York: McGraw-Hill Book Co, Inc, 1934.
- Ludemann, W. W., "Case Histories for All Pupils." *School and Society*, 43 509-10, April 11, 1936.
- Morgan, John J B, *The Psychology of the Unadjusted School Child* (Revised). New York The Macmillan Co, 1936
- Mort, Paul R, and Gates, Arthur I, *The Acceptable Uses of Achievement Tests*, Chapter I. New York Bureau of Publications, Teachers College, Columbia University, 1932.
- Olson, Willard C., "Diagnosis and Treatment of Behavior Disorders of Children." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XVIII, pp. 363-97. Bloomington, Ill Public School Publishing Co., 1935.
- Rivlin, Harry N., *Educating for Adjustment*. New York D. Appleton-Century Co, Inc, 1936.
- Seder, Margaret, *Introduction to Testing and the Use of Test Results*. New York Educational Records Bureau, July 1940
- Segel, David, *Nature and Use of the Cumulative Record* U. S. Office of Education Bulletin, 1938, No. 3. Washington, D. C Government Printing Office, 1938.
- Shaffer, Laurance Frederic, *The Psychology of Adjustment* Boston: Houghton Mifflin Co, 1936
- Sherman, Mandel, "Contributions to Education of Scientific Knowledge in Mental Hygiene" *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XXXV, pp. 435-43 Bloomington, Ill.: Public School Publishing Co, 1938
- Sherman, Mandel, *Mental Hygiene and Education*. New York. Longmans, Green and Co, 1934.
- Stogdill, Emily L, and Herndon, Audell, *Objective Personality Study A Workbook in Applied Mental Hygiene*. New York Longmans, Green and Co, 1939
- Symonds, Percival M, *Mental Hygiene of the School Child*. New York The Macmillan Co., 1934
- Thorpe, Louis P, *Personality and Life*. New York: Longmans, Green and Co, 1941
- Traxler, Arthur E, *Case-Study Procedures in Guidance* (Revised). Educational Records Supplementary Bulletin B New York. Educational Records Bureau, January 1940.

- Traxler, Arthur E, "A Cumulative Record Form for the Elementary School" *Elementary School Journal*, 40 45-54, September 1939.
- Traxler, Arthur E, *The Use of Tests and Rating Devices in the Appraisal of Personality* Educational Records Bulletin No 23 New York Educational Records Bureau, March 1938.
- Witty, Paul A, and Skinner, Charles E (Editors), *Mental Hygiene in Education*. New York Farrar and Rinehart, Inc., 1939.
- Wood, Ben D, "The Major Strategy of Guidance." *Educational Record*, 15 419-44, October 1934.

CHAPTER XIII

TESTS IN DIAGNOSIS AND REMEDIAL TEACHING

This chapter treats the following aspects of diagnostic testing and remedial teaching:

- a.* Meaning of diagnosis in education
- b.* Teacher responsibility for the diagnostic use of tests.
- c.* Exact diagnosis by means of educational tests.
- d.* Relation of remedial teaching to exact diagnosis.
- e.* Principles for the construction of remedial materials.
- f.* Concrete application of these principles to the construction and use of remedial materials.

I. THE PLACE OF DIAGNOSIS AND ANALYSIS

The Meaning of Diagnosis. Educational diagnosis implies the use of procedures, more or less technical in character, designed to locate specific learning and instructional difficulties, and if possible to determine their causes. For the medical expert, diagnosis means the careful and extensive observation of the patient under controlled conditions. It includes the use of professional instruments, such as the clinical thermometer, the stethoscope, the microscope, etc., which make observations more exact and more objective. It means the assembly of a complete case history of the background of the difficulty leading up to the present physical crisis. It is based on the examination and analysis of many similar cases, in order that common factors may be identified. For the teacher, diagnosis has many of the same implications, but unfortunately much of the exactness, objectivity, and precision of the medical diagnostician's instruments appear to be missing in the teacher's equipment. Only recently have objective measuring instruments capable of reasonably exact diagnosis become available to the pedagogical diagnostician. The well-prepared modern teacher now has at hand refined statistical techniques; exact and analytical diagnostic tests in different subjects; diagnostic

charts, devices for measuring aural acuity; and instruments for measuring visual acuity, eyedness, muscular imbalance in the eyes, binocular vision, and binocular fusion, and many other highly important qualities which may account for a pupil's lack of progress in many fields of learning. It is thus apparent that diagnosis in education is rapidly becoming accurate and scientific.

Diagnosis in Education. While diagnosis through measurement represents only a single phase of the work in education, it is nevertheless an exceedingly important aspect of it. The analysis and diagnosis of difficulties underlying educational accomplishments undoubtedly constitute the high point in the supervisory and instructional uses of educational tests. General impressions of strengths or weaknesses are supported or denied by test results. Deficiencies of a general nature are revealed and brought to light by general survey tests. Specific weaknesses, and to a certain extent causes of weaknesses, are identified by the use of properly selected diagnostic tests. Practically all of the more exact types of diagnostic procedures, such as the location of defects in speech, hearing, and vision, are dependent upon educational test results for their initial steps. These points will be discussed much more thoroughly in the later chapters on reading and language.

Analysis as the Basis of Diagnosis. The successful development of the many sets of habits which constitute the bulk of school learning depends upon the care with which the underlying and basic habits of the subjects themselves are recognized and utilized in the teaching. If, in arithmetic, a habit of responding correctly to the 100 basic addition facts is essential to success in addition, then this fact must be recognized and stressed. If it can be shown that, in addition to these basic skills, there are other related but somewhat higher skills which are fundamental to successful work, these also must be recognized and developed. If it can be shown that teaching a child to add consists not only in developing mastery of the 100 basic combinations, but also involves higher levels of skill, such as knowledge of the higher decade addition facts, bridging of the tens, control of the atten-

tion span, and carrying from one column to the next, the teacher's task is made much more obvious. Similarly, if it can be shown that silent reading comprehension is not a single isolated ability but a composite of many elements, such as knowledge of word meanings, ability to get meaning from sentences, ability to arrange thought units and sentence units into logically organized wholes, ability to find desired material quickly, etc., the teacher has a real basis for his instructional procedures. Language is another basic subject in which many delicately balanced skills are interwoven in an extremely complex manner. Here again the elements of achievement in the total process must be identified. Blind trust in general practice on the total skill must necessarily give way to the exact identification and discovery of the particular points of pupil weakness as a basis for special emphasis.

Good diagnosis must parallel exactly the processes of good teaching. Effective diagnostic materials in any school subject can be prepared only after the skills contributing to success in this field have been isolated and identified. Psychologically the reason for this is that on the whole the child learns to do what he practices and not something else. Remedial work, accordingly, can function only when the exact level at which pupil mastery breaks down has been located. Thus the analysis must be penetrating and the diagnosis must be exact.

Specific Nature of Diagnosis. Diagnosis must be more exact than broad statements of general functions. It is not enough to discover that a child is unable to read silently. The exact nature of his handicap must be revealed before it is possible to undertake a remedial program. The more specific the diagnostic information revealed, the more exactly the remedial material can be made to fit the need. To return to a frequently used illustration, it is found by diagnosis that the child is unable to add, but unless the exact point at which his mastery of addition breaks down can be determined by the diagnosis, teaching or remedial efforts are largely wasted. One of the outstanding reasons why more effective teaching and remedial work has not been done in certain

fields is that no adequate analysis of basic skills has been made. Concrete illustrations of this need are given in connection with a related discussion in this chapter.

Importance of the Diagnostic Interpretation of Test Results. There is nothing mysterious in standard or informal objective test results which produce immediate improvement of instruction in the classroom. Tests as such are incapable of improving instruction because of any inherent power of their own. Existing conditions are merely revealed by them, and these with the limitations implied by the validity and reliability of the particular instruments used. Remedial or corrective teaching is the result of deliberate constructive effort by the teacher after the particular points of weakness in the instruction of the pupils have been revealed by the tests. The ease, clearness, and directness with which these needs are revealed by the tests is a measure of their real educational value. Too few of existing standard tests are so constructed as to permit the interpretation of their results directly in terms of an effective remedial procedure. However, this seems to be no good reason for the failure of teachers to apply more directly the results of this work in testing to the improvement of their teaching practice. Just as the data revealed by the navigator's instruments require calculation and interpretation, so is it necessary to analyze test data carefully in order to make them the basis of a genuine remedial program. Nor is the teacher the only one to profit from a frank study of test scores. Pupils themselves will often take an active part in remedial enterprises when they are given the facts concerning their own performances on standard tests.

The interpretation of test scores and the planning of remedial procedures are the most difficult parts of the use of standard educational tests. Moreover, they are by far the most important parts. One of the greatest needs in education today is the provision for genuine diagnostic testing in all instructional fields, supplemented by valid remedial work designed to correct the weaknesses and defects of individual pupils as revealed by the tests. It has come to be recognized that the giving of a test with a study of the resulting scores will not in itself improve the classroom situation. This is

especially true if the test is of the general survey type. It is interesting to learn, as a result of using tests in the classroom, that a pupil or the entire class is below standard in the subject, but unless it is learned with some exactness what causes the low level of achievement, the testing program will do little if anything more than supply interesting information. Teachers and supervisors have a right to expect that something more constructive will be provided in exchange for the time required for classroom testing.

Exact Diagnosis the Basis for Remedial Work. Accurate diagnosis of class and individual pupil difficulties, coupled with application of specific remedy, is the heart of enlightened use of exact methods of teaching. The success of the remedial or corrective teaching depends upon the accuracy and detail with which the specific skills involved in the successful achievement in the subject are identified and isolated in the test. Tests of the general survey type, or tests which report single unanalyzed scores, cannot supply this information in sufficient detail. This is illustrated in the record of a single pupil given in Table XI.

This table shows the test scores of a fourth-grade pupil on four tests—two in arithmetic and two in language. Two of the tests are analytical and diagnostic, while the other two are excellent examples of tests reporting single unanalyzed scores. The tests used here are *Test D, Arithmetic*, of the 1941 *Iowa Basic Skills Test*, the *Compass Diagnostic Test in Arithmetic*, Test 3, *Multiplication of Whole Numbers*, the *Stanford Achievement Test (Language Usage Section)*, and *Test C, Language*, of the 1941 *Iowa Basic Skills Test*.

The pupil's relatively low score on the problem section of the *Iowa Basic Skills Arithmetic Test* and his score of 14 on the fundamental operations part of this test indicate that he is having some type of difficulty with his work in arithmetic. An examination of his test paper shows that he attempted only sixteen of the problems and missed four. Furthermore, not one of the problems he solved correctly called for the use of any but the very simplest multiplication. A glance at his work on the fundamental operations section of his paper at once makes this difficulty apparent. Of the nineteen examples he attempted, he solved fourteen

TABLE XI
RECORD OF TEST SCORES OF PUPIL A.L.C.
(Grade 4A, School L)

Tests and Parts	Pupil Scores	End-of-year Norm for Grade
<i>Iowa Basic Skills Tests, Elementary, Test D, Arithmetic</i> Fundamental Operations Problems	14 12	20 15
<i>Compass-Diagnostic Test in Arithmetic, Test 3, Multiplication of Whole Numbers</i> Part 1 Basic Multiplication Facts Part 2 Additions Used in Multi- plication in Parts 4, 5, 6 Part 3 Carrying in Addition Used in Parts 4, 5, 6 Part 4 Fundamentals in Multipli- cation Part 5 Checking Multiplication Part 6 Finding Errors Total	7 50 40 5 0 0 102	37 48 37 16 1 4 143
<i>Stanford Achievement Test, Intermediate, Test 3, Language Usage</i>	50	48
<i>Iowa Basic Skills Tests, Elementary, Test C, Language</i> I Punctuation II Capitalization III Usage IV Spelling V Sentence Usage Total	22 28 38 34 28 150	28 33 36 28 25 150

correctly, but no one of these involved long multiplication. Thus it is apparent that his difficulty lies somewhere in the field of multiplication.

The teacher at once gave him the *Compass Diagnostic Test in Arithmetic, Test 3, Multiplication of Whole Numbers*. His scores on this test are enlightening, as is shown in Table XI. Of the sampling of 61 basic multiplication facts which are used in the later sections of this test, this

pupil was able to give only seven. The end-of-the-year norm for his grade is 37. On Parts 2 and 3, which involve addition skills called for in long multiplication, his scores are above the grade norm. However, on Parts 4, 5, and 6, which deal with the specific application of the basic multiplication facts to multiplication examples, his scores are very low. This weakness makes for slow and inaccurate work in multiplication of the type called for in the general survey test. It is noticeable, however, that the survey test does not reveal in this specific manner the exact causes of his low score.

In the further diagnosis of this pupil's school achievement, two language tests were also used. The first, the *Language Usage* section of the *Stanford Achievement Test*, is a general survey of language usage. The pupil makes a score of 50 on this test, which is in excess of the norm for his grade. The administration of a more detailed analytical type of test, *Test C* of the 1941 *Iowa Basic Skills Test*, makes it very clear that while this pupil is very well equipped in word meanings, language usage, and spelling, he is below the expected levels of achievement for his grade in such abilities as punctuation and capitalization. Here again is an instructional situation which is not at all revealed by the general survey instrument.

Diagnosis as the Basis for Preventive Work. An examination of the number and types of skills identified as a result of the diagnostic methods discussed in the preceding section leads to a suggestion of a still more constructive use of analytical and diagnostic test results. Diagnosis as applied in education has taken on a meaning indicative of a breakdown in method, a failure of instructional techniques to function. Unquestionably one of the basic purposes of diagnosis is the location of weaknesses and the determination of their causes, but there is nothing in the method which precludes its use in the prevention of weaknesses through anticipation of their causes. Out of the knowledge gained through the use of diagnostic procedures should come the basis for preventive work of all types. It is quite noticeable that the major emphasis in the fields of dentistry and medicine is not on correction but on prevention. The existence of a weakness implies a failure at some point in the

program. The discovery of it should not be marked as important merely because it is then possible to correct it. The real importance in the discovery should lie rather in the prevention of its reappearance elsewhere under similar conditions.

Another illustration from the field of medicine may make this point somewhat more concrete. In every medical examination for diagnostic purposes, a complete analysis is made and an exact case record of all observations is kept. Out of the analysis of these records has come a better understanding of the causes and characteristics of certain types of human ailments. Out of this same type of analysis has also come the basis for much of the marvelous preventive work which characterizes modern medical science. In a similar way, accurate and detailed educational diagnosis may ultimately offer the basis for the development of a program of preventive work in education. For example, if, after diagnosing the addition of fractions in the fifth grade, it is found that the failure of pupils to reduce fractions in the answers is a common weakness, the obvious thing to do is to correct the defects at once, and then proceed to reconstruct the first instruction so that the following year the causes for this particular weakness may not operate so powerfully. Similarly, any weakness identified now should afford the basis for decisions calculated to reduce the probability of their recurrence in the future.

II. THE PLACE OF REMEDIAL INSTRUCTION

General Practice Exercises vs. Remedial Drill. There are in general two ways of maintaining a high level of pupil achievement in any subject after direct instruction has been discontinued. These are (1) systematic drill in general with no integral units of testing to discover breakdowns in pupil mastery, and (2) systematic remedial drill devices to fight forgetting, plus diagnostic testing to discover the exact causes of weaknesses when such weaknesses begin to cause poor work on review drills. The first method involves the systematic use of properly distributed general practice over the complete function. The second involves the periodical location of the specific defects of each pupil by means of diagnostic

tests and the immediate correction of these defects by the use of properly constructed remedial drill.

Unquestionably the latter is the more economical method of maintaining mastery of desired skills on the part of a pupil. It is obvious that general review is valuable at times, but just to review with no specific idea of what the review is to accomplish is too naive and hopeful to be effective. The program which coincides most closely with the experience of successful teachers and with a sound psychology of learning calls for the following steps in approximately the order indicated: (1) *teach*, (2) *review*, (3) *test for weaknesses* whenever they appear, and (4) *follow with remedial drill units* on the specific weaknesses revealed by the tests. It may be worth while to note that material so constructed as to be effective for remedial purposes is also sound to use for initial instruction. In fact, the chief distinction between good subject-matter content for initial teaching purposes and remedial drill purposes lies in *when* they are to be used. The most effective remedial drill for the pupil who does not have an adequate sight-meaning vocabulary for silent reading purposes is drill on the vocabulary he should have learned in the first place.

Necessity for Valid Drill for Each Identified Skill. Drill material of established validity must be provided for each specific skill which conditions achievement in the subject, if remedial work is to be effective. The validity of drill material depends to a large degree upon the accuracy and completeness with which the analysis of skills is made. Difficulties in subject-matter units which can only hazily be identified cannot be remedied except by mere good luck. Drills must closely parallel the skills which they are supposed to remedy. If mastery of a certain minimal vocabulary is essential to effective silent reading comprehension, then drill on those particular words which constitute special weakness should take precedence over other drill.

Perfect validity of drill material can be achieved only by taking a 100 percent sampling of all of the possible basic facts or skills in the particular field. Naturally this is impossible in certain cases, but it is nevertheless often possible to take such a large sampling that all of the most frequently

used facts are included. Subject-matter fields vary widely in the ways in which they lend themselves to sampling of this kind. In fields such as reading or language, a perfect sampling is almost impossible to obtain. On the other hand, many of the basic facts in arithmetic are so readily identified that they may be sampled 100 percent without difficulty. For example, a drill card for remedial (or initial instructional) purposes on the basic addition facts may sample the 100 basic facts, all of the basic addition facts there are. On the other hand, there are so many higher decade addition

SPECIMEN OF A TEACHER-MADE FRACTIONS REMEDIAL UNIT
CHOOSING COMMON DENOMINATORS

Directions — You can always get a common denominator by multiplying the denominators together. But sometimes smaller denominators can be used. Always use a small common denominator when you can. Practice on this exercise until you can choose denominators very quickly and accurately.

Write the denominator you would use in each example below :

Sample: $\frac{-}{2} + \frac{-}{4} + \frac{-}{3} = \frac{-}{12}$

$$\frac{-}{2} + \frac{-}{16} + \frac{-}{8} = \frac{-}{12} + \frac{-}{4} + \frac{-}{3} = \frac{-}{6} + \frac{-}{15} + \frac{-}{5} = \quad (3)$$

$$\frac{-}{4} + \frac{-}{6} + \frac{-}{2} = \frac{-}{6} + \frac{-}{6} + \frac{-}{9} = \frac{-}{5} + \frac{-}{3} + \frac{-}{3} = \quad (6)$$

$$\frac{-}{8} + \frac{-}{12} + \frac{-}{4} = \frac{-}{2} + \frac{-}{4} + \frac{-}{12} = \frac{-}{18} + \frac{-}{3} + \frac{-}{6} = \quad (9)$$

$$\frac{-}{10} + \frac{-}{4} + \frac{-}{5} = \frac{-}{5} + \frac{-}{4} + \frac{-}{2} = \frac{-}{6} + \frac{-}{12} + \frac{-}{8} = \quad (12)$$

$$\frac{-}{3} + \frac{-}{2} + \frac{-}{12} = \frac{-}{2} + \frac{-}{16} + \frac{-}{4} = \frac{-}{8} + \frac{-}{3} + \frac{-}{4} = \quad (15)$$

$$\frac{-}{24} + \frac{-}{2} + \frac{-}{8} = \frac{-}{6} + \frac{-}{5} + \frac{-}{10} = \frac{-}{12} + \frac{-}{6} + \frac{-}{9} = \quad (18)$$

$$\frac{-}{2} + \frac{-}{5} + \frac{-}{3} = \frac{-}{8} + \frac{-}{4} + \frac{-}{2} = \frac{-}{2} + \frac{-}{16} + \frac{-}{8} = \quad (21)$$

$$\frac{-}{5} + \frac{-}{2} + \frac{-}{5} = \frac{-}{12} + \frac{-}{8} + \frac{-}{4} = \frac{-}{6} + \frac{-}{3} + \frac{-}{16} = \quad (24)$$

facts (many of which have relatively slight social utility) that only a limited sampling can be taken. In practice it seems likely that a sampling of 450 such higher decade addition facts will cover most of the useful skills.

Drill material designed for remedial and corrective use strikes directly at the heart of the trouble and wastes no time on skills which need no practice. Remedial drill material in which careful control is kept over the distribution of practice on the basic skills is almost certain to be more effective than random exercises, even assuming in both cases that suitable motivation for improvement is provided. The drill will be most productive which most nearly provides a complete coverage of the skills of basic importance in the hierarchy of habits upon which successful achievement in the subject depends. Poorly organized drills may or may not deal with all possible weaknesses, but they are almost certain to waste time on skills which are not in need of drill. The validity of the drill depends upon the degree to which this sampling covers the basic or fundamental skills and the degree to which the exercises themselves actually develop the skills they purport to develop. There are a number of places in which this complex chain may break. The task of diagnostic and remedial treatment is to locate and repair quickly those links of the chain which have snapped under stress, or have rusted out through lack of use.

Necessity for Complete Coverage of Entire Area of the Skill. Correctly designed remedial material will not only parallel valid drill on the correct skills, but it will also cover all of the basic aspects of the skill. Furthermore it should acquaint the child with the most important variants of each situation. For example, in the development of skill in the subtraction of fractions, provision must be made for drill on the finding of common denominators, reduction of proper fractions, reduction of improper fractions, borrowing (or carrying) difficulty, etc.

Synthesis of Skills. Effective remedial material must not only cover in a valid manner all of the basic or underlying skills upon which achievement in the field depends, but it must provide a means for bringing about a gradual union of these component elements into the total function.

It is entirely possible that a mastery of the subsidiary skills involved might result in only a partial control of the end-product, if that goal were not reached by the gradual bringing together of each distinct skill in its relation to the whole process. This point may be illustrated by referring to the following tentative list of specific skills involved in mastery of the long division process. Experimentation and observation show that a breakdown in the total skill may take place at any one of the following points :

1. Knowledge of vocabulary of long division
2. The form of long division
3. The fundamental division facts
4. Carrying in long division
5. The multiplication involved in long division
6. The addition involved in long division
7. The subtraction involved in long division
8. The estimation of quotient figures, both apparent and non-apparent¹

From this it appears that there are at least eight basic skills and concepts which must be developed and brought into relation with each other if mastery of long division is to be attained. However, in order to provide adequately for the synthesis of these skills into the complete process, at least ten more remedial units are necessary.²

Functions of Time and Achievement Standards for Remedial Drill Material. The establishment of time and achievement norms for remedial material enables the child to discover for himself when he has reached a reasonable degree of speed in the work as a result of drill on a specific skill. It also affords a valuable motivating agency for the

¹ Quotients in long division are said to be *apparent* when the trial quotient resulting from dividing the left or first digit of the divisor into the first figure (first and second if the divisor will not go into the first figure) of the dividend is the actual

$\begin{array}{r} 31 \\ 12 \overline{)372} \end{array}$

or true quotient as in $12 \overline{)372}$. Here 1 goes into 3 exactly three times and the true first quotient figure is 3. Quotients are said to be *non-apparent* when the trial quotient

$\begin{array}{r} 19 \\ 17 \overline{)323} \end{array}$

obtained in a similar way is not the true quotient, as in $17 \overline{)323}$. Here the 1 goes into 3 three times, but the true first quotient figure is 1 and not 3 as it appears to be on first trial. These types of estimation of quotient figures account for a large part of the difficulty pupils encounter in long division.

² H A Greene, J W Studebaker, F B Knight, and G M Ruch, *Teachers' Manual Economy Remedial Exercises in Whole Numbers*. Scott, Foresman and Co., Chicago, 1927.

child and permits the evaluation of improvement in the skills in terms of units of time. The teacher will, of course, be interested especially in this latter aspect.

No particular significance is to be attached to the fact that many of the illustrations in the foregoing discussion of remedial materials have been taken from the field of arithmetic. Arithmetic happens to be a field in which rather complete identification of the basic skills has already been made. When other fields have been as carefully analyzed, remedial materials will be developed as fully as they are in the field of arithmetic.

Summary. This chapter is in a sense the point at which all of the foregoing discussion of the meaning and uses of tests as instruments for the improvement of classroom instruction has been aimed. The chapter should clinch in the mind of the teacher and student the often repeated fact that the underlying purpose of all testing is the accurate determination of class and individual pupil difficulties to the end that remedial instruction may follow. Of equal or even greater importance is the new angle which more exactly analytical and diagnostic testing gives to the preventive phases of instruction through the anticipation of causes of weakness or difficulty.

TOPICS FOR DISCUSSION

1. Expand the idea of the parallel between diagnosis and remedy in medicine and in education
2. What reasons can you advance for the failure to develop adequate diagnostic and remedial materials in all subject-matter fields?
3. What are the essential differences between remedial and preventive work in education?
4. Show how a critical analysis of subject matter is necessary to the development of diagnostic tests.
5. What are the essential differences between good diagnostic material and good remedial material?
6. Select a school subject and show how the basic skills may be identified (diagnosed) in a way similar to that suggested in the discussion in this chapter.
7. What is meant by synthesis of skills?
8. What are the effects of time and achievement standards on results in remedial materials?

9. To what extent are time and achievement standards fundamental to drill and corrective instruction in school subjects? Defend your position. (This may not be answered specifically in this chapter. It is food for thought.)
10. In a school field which you are likely to teach (your major or an important minor), suggest a number of specific skills which enter into successful work and parallel this with suggestions for remedial treatment.

SELECTED REFERENCES

- Baker, Harry J., and Traphagen, Virginia, *The Diagnosis and Treatment of Behavior Problem Children*. New York The Macmillan Co., 1935.
- Brueckner, Leo J., "Diagnosis and Remedial Teaching." *Encyclopedia of Educational Research*, pp. 392-99. New York. The Macmillan Co., 1941.
- Brueckner, Leo J., "General Methods Educational Diagnosis" *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XXVIII, pp. 333-40. Bloomington, Ill. Public School Publishing Co., 1938.
- Brueckner, Leo J., "The Principles of Development and Remedial Instruction." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XI, pp. 189-98. Bloomington, Ill. Public School Publishing Co., 1935.
- Brueckner, Leo J., "Techniques of Diagnosis." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter VIII, pp. 131-53. Bloomington, Ill. Public School Publishing Co., 1935.
- Brueckner, Leo J., and Melby, Ernest O., *Diagnostic and Remedial Teaching*. Boston Houghton Mifflin Co., 1931.
- Eurich, Alvin C., and Wrenn, C. Gilbert, "Appraisal of Student Characteristics and Needs" *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part I, Chapter II, pp. 31-87. Bloomington, Ill. Public School Publishing Co., 1938.
- Lee, J. Murray, *A Guide to Measurement in Secondary Schools*, Chapter VIII. New York D. Appleton-Century Co., Inc., 1936.
- Lincoln, Edward A., and Workman, Linwood L., *Testing and the Use of Test Results*, Chapters VIII-X. New York. The Macmillan Co., 1935.
- McCall, William A., *Measurement*, Chapter XXI. New York The Macmillan Co., 1939.
- Orleans, Jacob S., *Measurement in Education*, Chapter II. New York. Thomas Nelson and Sons, 1937.
- Pauli, Emanuel M., *Diagnostic Testing and Remedial Teaching*. New York D. C. Heath and Co., 1924.

- Ross, C. C., *Measurement in Today's Schools*, Chapter XIII New York . Prentice-Hall, Inc , 1941
- Tiegs, Ernest W , *Tests and Measurements for Teachers*, Chapter VI. Boston Houghton Mifflin Co , 1931.
- Tiegs, Ernest W , *Tests and Measurements in the Improvement of Learning*, Chapters II-III. Boston Houghton Mifflin Co , 1939
- Traxler, Arthur E , *The Use of Test Results in Diagnosis and Instruction in the Tool Subjects* (Revised). Educational Records Bulletin No. 18. New York Educational Records Bureau, January 1937.
- Tyler, Ralph W , "Characteristics of a Satisfactory Diagnosis." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter VI, pp. 95-111 Bloomington, Ill Public School Publishing Co , 1935
- Tyler, Ralph W., "Elements of Diagnosis." *Educational Diagnosis*. . Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter VII, pp. 113-29. Bloomington, Ill.. Public School Publishing Co , 1935
- Van Wagenen, M. J., *Educational Diagnosis and the Measurement of School Achievement*. New York The Macmullan Co., 1926.

CHAPTER XIV

MEASUREMENT AND REMEDIATION IN ARITHMETIC

This chapter summarizes the following points involved in the measurement of arithmetic skills and the diagnosis and remediation of pupil difficulties in arithmetic.

- a.* The curriculum in arithmetic.
- b.* Measurement of general achievement in arithmetic.
- c.* Diagnosis of basic arithmetic skills.
- d.* Diagnosis of problem-solving ability.
- e.* Types of remedial procedures in arithmetic.

Arithmetic more than any other subject has been aided by analyses of the basic skills involved. Diagnostic testing is common in arithmetic, and various types of remedial materials are available for use in following up the findings of diagnostic testing. Although problem solving is not as well served in these respects as are the basic arithmetic skills, a variety of materials are available even in that rather difficult area of performance. Because of the availability of many diagnostic and remedial materials and their consequent wide use, the emphasis in this chapter on arithmetic is somewhat more definitely centered upon diagnostic testing and remedial procedures than is the case for some other subject-matter fields.

I. COURSE CONTENT AND ORGANIZATION IN ARITHMETIC

At least three methods of approach to the question of what should be taught in arithmetic have been used: (1) social usage, (2) child usage, and (3) unit skills and problem types. The numerous studies of social usage which have been made since about 1910 have resulted in many modifications of the curriculum, chiefly through the elimination of such non-functional skills as cube root and cases in percentage, in order to adapt instruction more satisfactorily to the needs individuals encounter for arithmetic skills and abilities. The

child usage approach is similar, although it is more widely useful for the first two grades, during which informal number activities rather than formal instruction constitute the work in arithmetic. Both of these procedures are based on studies of social utility, which determine the types of arithmetic skills and abilities people actually have occasion to use in real life situations.

The third basis for determining the content and organization of the arithmetic curriculum supplements rather than conflicts with the first two. Based on the unit skills and types of problems important in the subject, it might be called the psychological approach.

Organization of Arithmetic Instruction. Of these three methods, the first was for a time responsible for a better selection of the types of arithmetic skills taught in the schools and the third has been chiefly responsible for the organization of subject matter for teaching and remedial purposes. Child usage studies have contributed mainly to instruction in the primary grades and to a certain extent in the activity school.

Application of the social utility theory eventually resulted in such a great reduction of content in arithmetic and in the time devoted to its teaching that specialists began to question the advisability of any further continuance of the reductionist trend and to recommend enrichment of the curriculum by the addition of new content. Knight expresses this attitude by stating that a better curriculum cannot result from a process of subtraction only, but that the determination of what must be added to the course of study is of comparable importance.¹

The drill method of teaching, which assumed that arithmetic facts and skills are largely unrelated and should consequently be taught in isolation, also came to be questioned seriously. Overman, in summarizing various studies of the degree to which arithmetic skills are transferred by pupils to numerical situations not previously encountered, concluded that instruction which stresses general rules, relationships,

¹ F. B. Knight, "Some Considerations of Method" *Report of the Society's Committee on Arithmetic* Twenty-Ninth Yearbook of the National Society for the Study of Education, Chapter IV, p. 149 Public School Publishing Co., Bloomington, Ill., 1930.

and methods of procedure is more effective than that which stresses isolated facts and skills.²

Brownell summarizes the "meaning" theory of arithmetic instruction, which has emerged during the last decade, by stating that, "within the 'meaning' theory there is absolutely no place for the view of arithmetic as a heterogeneous mass of unrelated elements to be trained through repetition. The 'meaning' theory conceives of arithmetic as a closely knit system of understandable ideas, principles, and processes. According to this theory, the test of learning is not mere mechanical facility in 'figuring.' The true test is an intelligent grasp upon number relations and the ability to deal with arithmetical situations with proper comprehension of their mathematical as well as their practical significance."³

The "meaning" approach does not abandon drill as a teaching device nor the results from psychological analyses of basic skills in the organization of instruction, but it attempts so to organize instruction that quantitative relationships become meaningful to the child. For example, Brownell points out that common fractions, decimal fractions, and percentage, commonly taught as different mathematical forms, are actually but three different ways of expressing the same ideas, and should be taught in that manner.⁴

List of Basic Arithmetic Skills. Arithmetic is one of the more definite tool subjects, and much of its subject matter is suitably organized for teaching purposes. For years it has been recognized that success in addition depended upon a mastery of the basic addition facts. The same may be said of each of the four fundamental processes with whole numbers. Teachers now recognize, however, that success in such work as long division is dependent upon a great many more skills than are involved in the mastery of the basic division facts. Long division calls for the accurate and effective use

² James R. Overman, "The Problem of Transfer in Arithmetic" *The Teaching of Arithmetic* Tenth Yearbook of the National Council of Teachers of Mathematics, pp 179-80 Bureau of Publications, Teachers College, Columbia University, New York, 1935

³ William A. Brownell, "Psychological Considerations in the Learning and the Teaching of Arithmetic" *The Teaching of Arithmetic* Tenth Yearbook of the National Council of Teachers of Mathematics, p 19 Bureau of Publications, Teachers College, Columbia University, New York, 1935

⁴ *Ibid*, p 26.

of skills in addition, multiplication, and subtraction, not to mention the skills which are usually recognized as being definitely division. Multiplication itself may involve the basic multiplication facts, the addition and multiplication involved in carrying in multiplication, and addition itself. A very complete catalogue of arithmetical skills selected for teaching, testing, and remedial purposes is the one on which the *Compass Diagnostic Tests in Arithmetic* are based. A summarization of this analysis is presented here to illustrate the extent to which such an analysis may be carried as well as to furnish a broad basis upon which to build diagnostic and remedial material in this field.

BASIC ARITHMETIC SKILLS, COMPASS DIAGNOSTIC TESTS⁵

I. Fundamental Processes with Whole Numbers

1. Basic Addition Facts
2. Basic Subtraction Facts
3. Basic Multiplication Facts
4. Basic Short Division Facts
5. Basic Vocabulary and Definitions of Arithmetic
6. Basic Rules of Arithmetic
7. Higher Decade Addition
8. Column Addition
9. Carrying in Column Addition
10. Harder Subtraction
11. Checking Errors in Subtraction
12. Borrowing or Carrying in Subtraction
13. Addition Used in Harder Multiplication
14. Carrying in Addition Used in Harder Multiplication
15. Complete Process of Multiplication
16. Short Division Involving Carrying
17. Multiplication, Addition, and Subtraction Used in Long Division
18. Complete Process of Long Division

II. Fundamental Processes* with Fractions and Whole Numbers

1. Changing Fractions to Equivalent Forms
2. Finding Common Denominators
3. Reducing Fractions
4. Addition of Fractions and Mixed Numbers
5. Expressing Mixed Numbers as Improper Fractions
6. Fundamentals of Subtraction of Fractions

⁵ Adapted from G M Ruch, F B Knight, H A Greene, and J W Studebaker, *Manual of Directions for Compass Diagnostic Tests in Arithmetic*, pp 9-12 Scott, Foresman and Co, Chicago, 1925

7. Reduction of Mixed Numbers
8. Cancellation in the Multiplication of Fractions
9. Reduction of Fractions and Mixed Numbers to Best Form in Answer
10. Multiplication of Fractions
11. Cancellation in Division of Fractions
12. Changing from Multiplication to Division Form
13. Fundamentals of Division of Fractions
- III. Fundamental Processes with Decimals
 1. Notation of Decimals
 2. Changing Fractions and Mixed Numbers to Decimal Form
 3. Changing Decimals to Fractions and Mixed Numbers
 4. Fundamentals of Addition of Decimals
 5. Fundamentals of Subtraction of Decimals
 6. Place Values in Decimals
 7. Pointing off in Multiplication of Decimals
 8. Dividing Decimals by Pointing off
 9. Location of Decimal Points in Division
 10. Changing Remainders to Decimal Form
 11. Fundamentals of Division of Decimals
- IV. Fundamental Processes with Denominate Numbers
 1. Knowledge of Tables of Measure
 2. Reducing in Denominate Numbers
 3. Borrowing in Denominate Numbers
 4. Addition of Denominate Numbers
 5. Subtraction of Denominate Numbers
 6. Multiplication of Denominate Numbers
 7. Division of Denominate Numbers
- V. Mensuration
 1. Vocabulary of Mensuration
 2. Mensuration of Plane Surfaces
 3. Mensuration of Solids
 4. Finding Areas and Volumes
 5. Formulas Used in Mensuration
- VI. Percentage
 1. Fractional and Percent Relations
 2. Decimal and Percent Relations
 3. Expressing Areas in Percents
 4. Fundamentals of Work in Percentage
- VII. Interest
 1. Vocabulary of Interest
 2. Business Forms
 3. Budgets
 4. Computation of Interest
 5. Computation of Discount
 6. Use of Interest Tables

VIII. Problem Solving

1. Mastery of Fundamental Processes
2. Comprehension of Material Read in Problem
3. Knowledge of What Is Given in the Problem
4. Knowledge of What Is Called for in the Problem
5. Probable Answer to the Problem
6. Knowledge of Proper Processes to Use in Solving
7. Knowledge of Proper Order of these Processes
8. Recognition of the Correct Solution

II. MEASUREMENT OF GENERAL ACHIEVEMENT IN
ARITHMETIC

A comprehensive understanding of standardized tests and their nature and use is best attained by an examination of sample tests and, if possible, such accompanying materials as manuals of directions, scoring keys, and pupil record forms, or, preferably, the actual use of one or more tests in the classroom. Therefore, in this and other similar chapters the authors have chosen to present only a few sample items from various tests to illustrate the application of different objective testing methods to various subject-matter fields. Although practice varies according to the subject-matter field in this respect, such representative items will quite largely take the place of discussions of specific standardized tests. To conserve space, directions to the pupils are not given for the sample items except in instances of unusually complex item forms. The student should be sufficiently familiar with various item forms and their modifications, as presented in Chapter VIII, that his interpretation of the samples given here should not be affected by the absence of such directions.

It is believed that the presentation of sample items with brief comments will serve two valuable purposes (1) familiarize the student with information concerning standardized testing methods and major item techniques in arithmetic, and (2) furnish him with suggestions concerning some of the methods he may very well apply in constructing tests for use with his own classes.

Standardized Testing in Computational Skills. Computational skills are most often tested by an item type of simple recall form, although multiple-choice items are sometimes used. Such item types can be used with any com-

bination of the four fundamental operations—addition, subtraction, multiplication, and division—and the four types of numbers—whole and mixed numbers, fractions, and decimals. Some tests classify all items of a type together, while others use the “omnibus” arrangement of mixed order for the various operations and types of numbers.

Simple Recall Items. Although these items are of simple recall form, it is by means of performing certain calculations rather than as recall that a pupil obtains the answers. Directions are usually given to the pupil concerning the form of answer desired, e.g., mixed numbers reduced to whole numbers and fractions, fractions reduced to lowest terms. Definite rules are also usually provided in order to objectify the scoring of a type of performance which is often viewed by different teachers according to very different standards. Credit is ordinarily given only for answers which are entirely correct.

Sample A.⁶

1. Subtract	2. Subtract	3. Multiply	4. Divide	5. Add	1.
$\begin{array}{r} 46 \\ -25 \\ \hline \end{array}$	$\begin{array}{r} 658 \\ -101 \\ \hline \end{array}$	$\begin{array}{r} 601 \\ \times 5 \\ \hline \end{array}$	$42 \div 7 =$	$\begin{array}{r} 13 \\ 21 \\ +42 \\ \hline \end{array}$	2.
					3.
					4.
					5.

Sample B.⁷

1 Add $\begin{array}{r} 8 \\ 6 \\ \hline \end{array}$	6 Subtract $\begin{array}{r} 9 \\ 3 \\ \hline \end{array}$	11 Add $\begin{array}{r} 174 \\ 937 \\ 425 \\ 801 \\ \hline \end{array}$	16 Divide $\begin{array}{r} \boxed{} \\ 4 \overline{) 12} \end{array}$
-----------------------------------------------------------------	----------------------------------------------------------------------	------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------

Multiple-Choice Items. Multiple-choice items require the pupils to perform the calculations in order to determine which is the correct answer, although there is usually no re-

⁶ Martha Kellogg, L. J. Brueckner, and M. J. Van Wagenen, *Analytical Scales of Attainment Arithmetic*, Grades 3-6. Published by Educational Test Bureau, 1933.

⁷ H. F. Spitzer, *Iowa Every-Pupil Tests of Basic Skills Test D, Basic Arithmetic Skills*, Elementary. Published by Houghton Mifflin Co., 1940.

quirement that the pupil put down the work by which he obtained the answer. Some pupils might obtain the answers by mental computation and others by putting down only a skeleton of their computations.

Sample C.⁸

$\begin{array}{r} (41) \\ 533 \\ \times 4 \\ \hline \end{array}$	$\begin{array}{r} (42) \\ 400 \\ \times 3 \\ \hline \end{array}$	$\begin{array}{r} (43) \\ 509 \\ \times 8 \\ \hline \end{array}$	<p>41. Ans. ^a537 ^b2132 ^c529 ^d133$\frac{1}{4}$</p> <p>42. Ans. ^a403 ^b133$\frac{1}{3}$ ^c1200 ^d397</p> <p>43. Ans. ^a501 ^b517 ^c63$\frac{1}{2}$ ^d4072</p>
------------------------------------------------------------------	------------------------------------------------------------------	------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Some evidence resulting from experimenting with the multiple-choice technique of testing in arithmetic indicates a considerable increase in the validity of this type of item by the addition of another answer space for the response "Correct Answer Not Given." A specimen of this type of exercise in which the correct answer is not given is shown below.

Sample D.⁹

$\begin{array}{r} \text{Add: } 736 \\ 618 \\ 422 \\ 907 \\ \hline \end{array}$	<p>(1) 1683</p> <p>(2) 2693</p> <p>(3) 438</p> <p>(4) 3795</p> <p>(5) Correct answer not given</p>
--------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------

Standardized Testing in Problem Solving. Standardized tests in problem solving are most frequently set up either in simple recall or in multiple-choice form. The four examples given below are sufficient to illustrate the testing method because of the similarity of problem solving items in different tests.

Simple Recall Items. Simple recall items in this situation require solutions of the problems, rather than recall in the usual sense of that word, in obtaining the answers. Scoring of responses is practically always on an all-or-none basis,

⁸ Ernest W. Tieg and Willis W. Clark, *Progressive Arithmetic Test*, Intermediate Published by California Test Bureau, 1939

⁹ Adapted from an experimental test devised by G. W. Maxwell. See G. W. Maxwell, *The Use of Multiple Choice Items in Measuring Achievement in Arithmetic*. Master's thesis, University of Iowa, Iowa City, 1940

for no credit is given unless the answer is correct. Only one illustration of this item type is shown.

Sample E.¹⁰

1. I bought an apple for 4 cents, a bowl of soup for 8 cents, and a cookie for 2 cents. All of the food cost how many cents?
2. John has 6 cents and wants to buy a ball that costs 15 cents. How many more cents does he need to buy the ball?

Multiple-Choice Items. The multiple-choice item in problem solving also usually requires the solution of the problem in order to determine which of the alternative answers is the correct one. However, some items require only an indication of the information necessary in a problem situation.

Sample F.¹¹

41. In a class room there were 7 rows of desks with 7 desks in each row. Four desks were removed from the room. How many desks were left?
Ans ^a52 ^b45 ^c38 ^d9
42. Henry bought a used automobile for \$55.00. He paid \$10.00 down and is to pay the rest in nine equal payments. How much will each payment be?
Ans ^a\$10.00 ^b\$4.00 ^c\$5.00 ^d\$9.00

Sample G.¹²

2. You have paid five cents for five pieces of candy. You want to sell one piece of candy for what it costs you. **BEFORE SELLING THE CANDY, YOU SHOULD FIND OUT**
 - d how many pieces of candy you can buy for ten cents
 - c how much you paid for one piece of candy
 - f how many cents there are in a nickel

Sample H.¹³

- A rectangular field is 26 rods long and 18 rods wide. How many rods of wire netting are required to enclose it?
- (1) 468
 - (2) 108
 - (3) 668
 - (4) 44
 - (5) Correct answer not given

¹⁰ Gertrude H. Hildreth, *Arithmetic Achievement Tests*, Grades 2 to 6. Published by Bureau of Publications, Teachers College, Columbia University, 1935.

¹¹ Tiegs and Clark, *op cit*.

¹² Robert K. Speer and Samuel Smith, *National Achievement Tests Arithmetic Reasoning*, Grades 3 to 8. Published by Acorn Publishing Co., 1938.

¹³ Maxwell, *op cit*.

Standardized Testing of Basic Concepts. Several of the modern arithmetic tests include sections for the measurement of arithmetic vocabulary, meaning of symbols, quantitative relationships, and other knowledges distinct from those directly involved in computation and problem solving. A few illustrations of such items, which are mostly of the multiple-choice type, are given below.

Sample I.¹⁴

- | | | | | | | |
|-------------|--------------|--------------|------------|-------------|-------------|--------|
| 1. double | 1 once | 2 different | 3 make | 4 twice | 5 construct | 1. ... |
| 2. clock | 1 time piece | 2 number | 3 distance | 4 length | 5 cost | 2. . |
| 3. subtract | 1 add | 2. take away | 3 divide | 4. increase | 5 break | 3. . |

Sample J.¹⁵

- | | | |
|--------------------------|-----------------------|--------------------------|
| 22. % means | ¹ subtract | per cent |
| | ³ dram | ⁴ dollar |
| 23. $\sqrt{\quad}$ means | ¹ add | ² ounce |
| | ³ interest | ⁴ square root |

III. DIAGNOSTIC TESTING IN ARITHMETIC SKILLS

Tests as such are incapable of improving instruction directly. Existing conditions are merely revealed by them, and it is worthy of note that these conditions are revealed only within the limits of the validity and the reliability of the particular tests used. The importance of using tests which are themselves based upon a sufficiently detailed analysis of the skills required for successful achievement in the field to permit the application of definite remedial procedures can hardly be over-emphasized. Remedial teaching is the result of deliberate instructional effort on the part of the teacher after the particular points of weakness of the pupils have been revealed. The accuracy with which these needs are revealed by the device used is the best measure of its value to the classroom teacher. Specific criteria for the evaluation of test materials are discussed in Chapter IV. Suggested criteria for the evaluation of available material designed for remedial instruction purposes are given in Chapter XIII. The same principles are certain to be useful to the student or

¹⁴ Kellogg, Brueckner, and Van Wagenen, op cit

¹⁵ Tiegs and Clark, op. cit.

the teacher who is interested in the preparation of his own remedial materials.

Scope of Diagnostic Testing in Arithmetic. It is not just chance that diagnostic tests have been developed in subject-matter fields where the aims are clean cut and in which the basic skills conditioning achievement have been analyzed carefully. Nor is it chance that the blanket purposes of certain other subject-matter fields, as expressed in courses of study and textbooks, have left the teacher groping vaguely for tangible goals and effective instructional methods. The order of development is clear: first, there must be a specific statement of aims lying back of the subject matter, second, a detailed analysis must be made of the basic skills upon which ultimate achievement depends; and third, material designed to give mastery of these skills must be prepared.

Some progress has been made in the diagnosis of pupil defects in the field of arithmetic. This is possible because the aims of arithmetic are quite clearly stated, which in turn permits a rather detailed analysis of the underlying skills. As soon as it became known, for example, that the ability to do a certain type of column addition depends upon the pupil's knowledge of approximately 450 higher decade addition facts, it was possible not only to locate difficulties in teaching this material as such, but also to furnish the teacher with specific aids in teaching it. The reason why similar material is not available in such fields as geography, history, and science is that the aims of instruction in these fields have not yet become sufficiently crystallized to permit the type of analysis to which arithmetic has been subjected.¹⁶

Diagnostic Tests in Basic Arithmetic Skills. Among the diagnostic tests are three useful series, each representing a rather specific point of view in diagnosis. Although none

¹⁶ It should probably be pointed out here that very likely there are certain subject-matter fields in which this type of crystallization of aims will not and should not take place. This is undoubtedly true of certain subjects such as social studies and natural and physical sciences, in which changes in content are appearing so rapidly and in which there are certain points on which general agreement cannot be expected. Here the diagnosis will remain for some time in general terms, such as the ability to read or the ability to work the mathematics involved in certain science fields. Here also remedy will be largely in terms of bringing about a more adequate mastery of certain definite materials.

of these tests is of recent copyright, not a great deal of work has been done during recent years on diagnostic testing in arithmetic. The *Buswell-John Diagnostic Chart for Fundamental Processes in Arithmetic* is designed for individual diagnostic work. It consists of a diagnostic chart and a test sheet on which the pupil does his work aloud in the presence of the teacher. On the diagnostic chart, which is for the teacher's use, are listed the most frequent faulty habits of work and causes of error in the particular arithmetic process under diagnosis. The pupil is given the work sheet and instructed to work on each of the exercises, doing his work aloud. In this way the teacher is able to discover the pupil's method of work and check the major causes of his difficulty.

The *Brueckner Diagnostic Tests*, which cover whole numbers, fractions, decimals, and percentage, are really inventory exercises which make it possible for a sufficiently critical and analytical teacher to check through the pupil's work and discover the apparent causes of difficulty. The diagnosis thus becomes a matter of working out an individual analytical record for each child.

The *Compass Diagnostic Tests* represent the third of the approaches to diagnosis. This series consists of twenty tests covering the fundamental processes with whole numbers, fractions, decimals, percentage, arithmetical definitions and concepts, business forms, mensuration, and problem analysis and problem solving. The tests are essentially analytical in structure, the total process in each case being torn down one step at a time as a basis for the identification of the causes of weakness. These tests are designed for group measurement and diagnosis. It is probable that no diagnostic test in any field is capable of indicating precisely *why* a skill breaks down, but there is not much more certainty that the teacher's interpretation of *why* the pupils encountered difficulty will be much more exact. The list of skills enumerated in the outline on pages 307 to 309 gives a very definite idea of the wide range of ability covered by these tests.

EXCERPT FROM COMPASS DIAGNOSTIC TESTS

PROBLEMS	PART 1—COMPREHENSION	PART 2—WHAT IS GIVEN
<p>Read each problem below. Then work across the two facing pages to the right, doing all the Parts for one problem before going to the next. Do not go back and work on a Part after you have completed the one following.</p> <p>Read the Sample below.</p>	<p>Put a cross (X) on the line before the one statement below which is true for each problem.</p>	<p>Put a cross (X) on the line before every statement below which tells a fact given in the problem.</p>
<p>Remember Work across the page to the right.</p>		
<p>[Read the problem]</p> <p>Problem 1</p> <p>Our baseball team played 7 games this summer. We lost 2 and tied none. How many games did we win? →</p>	<p>[Check true statement]</p> <p>— Team won all games played</p> <p>— Team lost all games played</p> <p>— Team won more than it lost.</p> <p>— Team lost half of the games played.</p> <p>— Team won about half of the games played →</p>	<p>[Check what is given]</p> <p>— Number of games played</p> <p>— Number of boys on team.</p> <p>— Number of games tied</p> <p>— Number of games won.</p> <p>— Number of games lost →</p>

IV. TESTING OF PROBLEM-SOLVING ABILITY.

Meaning of Problem Solving. The social importance of problem solving in arithmetic is far in excess of the attention given to its analysis and measurement up to the present time. It is a complicated field which is almost certainly highly related to general intelligence. Naturally an attempt to analyze and identify the underlying skills meets with considerable difficulty. Thus far five fundamental steps in problem solving, closely paralleling the steps in the thinking process outlined by Dewey,¹⁷ have been identified. These steps afford practically the only workable basis for an attack upon problem-solving difficulties.

The first step in the solution of verbal problems demands a complete understanding of the elements and processes which are involved or implied. This is *comprehension*. This in itself involves many factors, such as rate of reading, vocabulary difficulties, reading of numerals, and problem organization, as well as complexity in terms of the number and order of the arithmetical processes involved. Underlying all of these is, of course, the ability of the child to hold

¹⁷ John Dewey, *How We Think* D C Heath and Co, Boston, 1910

IN ARITHMETIC, TEST XVII, PROBLEM ANALYSIS¹⁸

PART 3—WHAT IS CALLED FOR	PART 4—PROBABLE ANSWER	PART 5—CORRECT SOLUTION
Put a cross (x) on the line before the one statement below which tells what is called for in the problem.	Put a cross (x) on the line before the one statement below which gives the nearest probable answer to the problem. Do not take time to work the problem.	Put a cross (x) on the line before the one correct solution given for each problem. Figure in the margin if you want to.
Remember Work across the page to the right.		
[Check what is called for] ___ Number of games lost ___ Number of boys on team ___ Number of games won. ___ Number of games where score was tied. ___ Number of games played →	[Check probable answer] ___ 9 boys ___ 6 schools ___ About 7 boys ___ 9 games ___ About 5 games. →	[Check correct solution] ___ $7+2=9$ ___ $7-2=5$ ___ $7+3=3\frac{1}{2}$ ___ $7 \times 2=14$ ___ $7 \times 2=14, 14-9=5$ Now Start Problem 2.

the various facts and conditions in his mind long enough to analyze and organize them. This process of *analysis and organization* constitutes a second important step. The unnecessary facts or implications are discarded and only the significant data are retained. The third step in practice is actually a part of the second, for the *recognition of the process* involved is really a part of analysis. From this the worker moves straight to the fourth step, *solution*, where he applies to a specific situation his knowledge of the fundamental tools of number. In his earlier practice he has learned how to perform certain simple arithmetical skills. Now he learns when to apply them. The next and final step in the process is *verification*, which may be either a rough checking by the estimation of the probable answer to the problem, or an actual re-calculating and rechecking of the processes involved.

Diagnostic Tests of Problem-Solving Ability. Many children naturally fail to proceed in the solution of problems in so orderly a fashion as is indicated in this discussion,

¹⁸ G. M. Ruth, F. B. Knight, H. A. Greene, and J. W. Studebaker, *Compass Diagnostic Tests in Arithmetic, Test XVII, Problem Analysis*. Published by Scott, Foresman and Co, 1925

although it would be economical for them to do so. Oftentimes imperfect work (though finally successful) means using unnecessary steps, and spending useless energy in doing essential steps in an ineffective order.

So far in the development of diagnostic instruments for the identification of the skills involved in problem solving only a rough sampling of these skills has been approximated. The only two attempts of any consequence in this field are those represented by *Stevenson's Problem Analysis Tests* and the *Compass Diagnostic Tests on Problem Solving*.

Each of the *Compass Tests* (Tests XVII and XVIII) presents 15 problems on which exercises covering (1) Comprehension, (2) What is Given, (3) What is Called For, (4) Probable Answer, and (5) Correct Solution, are given. In both of the tests an attempt was made to parallel as closely as possible the foregoing introspective analysis of the skills involved, but only for purposes of identification. The arrangement of the exercises in the *Compass Diagnostic Tests on Problem Solving* (Test XVII, for grades 5 and 6) is illustrated in the accompanying reproduction of Problem 1.¹⁰

V. REMEDIAL INSTRUCTION IN ARITHMETIC

Remedial Materials in Arithmetic Fundamentals. The reader should keep in mind at all times that pupils do not fail in a vague, general sense, nor do they need remedial work of a vague and general type. Pupils' errors and failures are specific. The more exactly they can be located, the more promptly they can be removed. Diagnostic tests based upon a satisfactory analysis of the skills which are essential to pupil mastery are for the purpose of locating these specific breakdowns.

Remedial exercises incorporating most of the desirable characteristics of such material can be developed by the classroom teacher, or can be secured in commercial form in certain school subjects. The preparation and use of such

¹⁰ In the test itself the exercises are arranged so that the pupil works across the page to the right, answering all of the exercises dealing with a given problem before taking up another problem. In this reproduction the size of the page makes this type of arrangement impossible.

material to supplement available instructional devices should serve to increase greatly the efficiency of teaching. It should be remembered that the most effective use of remedial material will follow the careful diagnosis of individual pupil difficulties by means of tests prepared for the purpose, and that the use of the tests without the accompanying remedial program is equally futile.

An examination of available practice and drill material in the field of arithmetic reveals two somewhat distinctive types and uses of such material. General practice exercises designed to simplify the first learning and to aid in maintaining a general mastery of the skill are numerous and varying in type. They range from practice cards designed for repeated use to cheap practice tablets and comprehensive workbooks all designed for drill and maintenance purposes. The arithmetic drill devices which have been constructed particularly for remedial purposes are not so numerous.

In the field of commercial materials for remedial work in arithmetic the *Economy Remedial Exercises* may be used as illustrations of a program for meeting the practical problems of remedial work in this subject. Some idea of the organization of this material within a given field of arithmetic may be gained from the accompanying table, in which the units of remedial drill designed to correct difficulties in the manipulation of whole numbers are shown in their relation to the basic skills to be developed.

Problem-Solving Exercises. The development of skill in the solution of verbal problems is one of the more important goals of instruction in arithmetic. It is at the same time one of the more difficult skills to develop because of its complexity. The remedial aspect of the field of problem solving largely remains to be developed, although some definite remedial material is available in the *Economy Problem-Solving Exercises*. Since a complete sampling of the complex mass of skills involved in problem solving cannot be made, only a few of the most important ones are included in the proposed remedial material. Practice on the silent reading comprehension of verbal problems of varying degrees of complexity is given. Practice on the selection of the items given in the problem essential to its solution is

TABLE XII

SCOPE OF DRILL UNITS IN WHOLE NUMBERS, ECONOMY REMEDIAL EXERCISES ²⁰

For weaknesses in these Basic Skills	Use these types of Remedial Exercise Units
Basic Addition Facts	100 Addition Facts
Basic Subtraction Facts	100 Subtraction facts; easy combinations, no carrying or borrowing
Basic Multiplication Facts	100 Multiplication Facts
Basic Division Facts	90 Division Facts
Higher Decade Addition	450 Higher Decade Addition Facts
Column Addition	450 Higher Decade Addition Facts
Carrying in Column Addition	Introducing easy carrying
Harder Subtraction	Harder subtraction combinations introducing borrowing or carrying
Addition Used in Harder Multiplication	Higher Decade Addition Facts
Carrying in addition used in Harder Multiplication	The 360 multiplication and addition combinations used in carrying in multiplication
Complete Process of Multiplication	Units introducing one-, two-, three-, and four-figure multipliers, and zero difficulties
Short Division with Carrying	The 360 short division combinations involving carrying
Addition, Subtraction, and Multiplication involved in Long Division	Previous units in these fields
Estimating Quotients	Units introducing estimation of apparent and non-apparent quotients
Complete Process of Long Division	Units introducing the complete process of long division one step at a time

²⁰ Adapted from H. A. Greene, J. W. Studebaker, F. B. Knight, and G. M. Ruch, *Economy Remedial Exercises* Published by Scott, Foresman and Co., 1927

afforded by these exercises. Skill in comprehending the real problem setting by determining what is called for in the problem is also developed by much practice on this type of exercises. Practice on the basic skill of choosing the correct process or combination of processes in the more complex problems is also given. Skill in the verification of the solution or the estimation of the most probable answer to the problem is developed by a special set of exercises. The complete set of problem exercises can be utilized as a means of unifying the skills involved in the complete process of problem solving. Table XIII shows in compact form the relationship between the steps involved in problem solving, the factors underlying success in problem solving, and the types of drill provided by the *Economy Problem-Solving Exercises*.

In this manner at least six of the basic skills of problem solving are provided with remedial drill. The validity of the drill depends upon the degree to which this sampling covers the basic or fundamental skills and the degree to which the exercises themselves actually develop the skills they purport to develop. There are a number of places in which this complex chain may break. The task of diagnostic and remedial treatment is to locate and repair quickly those links of the chain which have snapped under stress, or have rusted out through lack of use.

Two observations may serve as an ending for this chapter. First, while specific drill on some skill all by itself is often quite important, it must be accompanied by, if not preceded by, understanding of the total situation. Satisfactory performance on an isolated skill is not always matched by similar performance on the same skill when it operates in a more complex situation. Thus $9 + 4$ may be an easy combination by itself, but it may not click at all when presented as $7 + 2 + 4$, where the 9 is unseen. From this it follows that after attention is paid to a specific breakdown the skill should also be practiced in the most complex situation in which it appears.

A second caution deals with a matter of policy. The need for remedial work of any kind and in any subject implies a failure at some point in the initial learning. Remedial work

TABLE XIII

ANALYSIS OF PROBLEM SOLVING, ECONOMY PROBLEM-SOLVING EXERCISES ²¹

Steps in Problem Solving	Factors Underlying Problem Solving	Types of Drill Provided
Comprehension	Vocabulary Ability to Read Numerals. Ability to Read Rapidly. Ability to Comprehend. <i>a</i> Follow directions <i>b</i> . Make generalizations. <i>c</i> Select potent elements. <i>d</i> Discard irrelevancies. <i>e</i> Determine problem setting as a unit <i>f</i> Determine the outcome of the problem. <i>g</i> Grasp the significance of problem cues.	Multiple Choice Comprehension Exercises
Analysis and Organization	Selection of Potent Elements Selection of Processes Involved. Determining What the Problem Calls For Determining What is Given in the Problem. Determining the Process Relationships	What is Given Process Analysis What is Called For Problem Relationships
Recognition	Choice of Procedure Determining Problem Conditions Determining the Purpose of the Problems Determining Relevant Elements	Process Analysis What is Called For What is Given
Solution	Selection of Process. Organization of Processes in Order Knowledge and Application of Combinations Problem Relationships.	Process Analysis Problem Scales
Verification	Probable Form of Answer. Probable Magnitude of Answer.	Probable Answer

²¹ Adapted from H A Greene, J W Studebaker, F. B Knight, and G M. Ruch, *Economy Problem-Solving Exercises* Published by Scott, Foresman and Co, 1928.

should be reduced as much as possible by making the first learning effective, by adequate review devices, and by the proper grading of pupils. A teacher should never be proud of the amount of remedial work he must do. However, he may be proud of his ability to direct it well when need for it arises. Obviously preventive work based upon understanding is better teaching than remedial work.

TOPICS FOR DISCUSSION

1. What accounts for the fact that the field of arithmetic has been subjected to rather extensive analysis and intensive measurement?
2. Identify some of the more important of the specific skills in arithmetic which appear to lend themselves to measurement and remedial treatment.
3. To what extent does it appear justifiable to depend upon transfer to aid in the learning of arithmetic skills?
4. Illustrate some of the standardized test techniques which have been used in the measurement of general skills, problem-solving ability, and basic arithmetic concepts.
5. Show how the basic steps in problem solving closely parallel the steps in the thinking process.
6. Describe the techniques proposed for the analysis of problem-solving abilities.
7. What are the chief differences in the principles underlying the three types of diagnostic tests in basic arithmetic skills?
8. Outline a survey, diagnostic, and remedial program in arithmetic, indicating where possible your first and second choices of material. Give reasons for your choices.

SELECTED REFERENCES

- Arithmetic in General Education* Sixteenth Yearbook of the National Council of Teachers of Mathematics New York Bureau of Publications, Teachers College, Columbia University, 1941.
- Beattie, Louise, "Standardized Tests in Arithmetic" *Educational Method*, 16 175-76, January 1937.
- Broom, M. E., *Educational Measurements in the Elementary School*, Chapter VIII. New York McGraw-Hill Book Co., Inc., 1939.
- Brownell, William A., "Remedial Cases in Arithmetic." *Peabody Journal of Education*, 7 100-7, September 1929.
- Bruceckner, Leo J., "Diagnosis in Arithmetic" *Educational Diagnosis*, Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XIV, pp. 269-302 Bloomington, Ill. Public School Publishing Co., 1935.

- Brueckner, Leo J, *Diagnostic and Remedial Teaching in Arithmetic*. Philadelphia The John C. Winston Co., 1930.
- Brueckner, Leo J, "Significant Trends in Research in Diagnosis in Arithmetic." *Journal of Educational Research*, 33 460-62, February 1940
- Brueckner, Leo J, and Melby, Ernest O, *Diagnostic and Remedial Teaching*, Chapter VII. Boston Houghton Mifflin Co., 1931
- Buswell, G T, "Contributions of Research to Special Methods Elementary School Mathematics" *The Scientific Movement in Education* Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, pp 123-28. Bloomington, Ill Public School Publishing Co, 1938
- Buswell, G T, and John, Lenore, *Diagnostic Studies in Arithmetic*. Chicago University of Chicago Press, 1926.
- Gilliland, A R, Jordan, R H, and Freeman, Frank S, *Educational Measurements and the Class-Room Teacher* (Revised Edition), Chapter IX New York The Century Co, 1931
- Greene, Charles E, and Buswell, G T., "Testing, Diagnosis, and Remedial Work in Arithmetic" *Report of the Society's Committee on Arithmetic* Twenty-Ninth Yearbook of the National Society for the Study of Education, Chapter V, pp 269-316. Bloomington, Ill Public School Publishing Co, 1930.
- Greene, Harry A, "A Critique of Remedial and Drill Materials in Arithmetic" *Journal of Educational Research*, 21 262-76, April 1930
- Klapper, Paul, *The Teaching of Arithmetic*. New York D. Appleton-Century Co, Inc, 1934
- Knight, F B, "Crucial Questions on Arithmetic Testing" *Chicago Schools Journal*, 11 4-9, 46-48, September and October 1928
- Knight, F. B (Chairman), *Report of the Society's Committee on Arithmetic* Twenty-Ninth Yearbook of the National Society for the Study of Education Bloomington, Ill Public School Publishing Co, 1930
- Madsen, I. N., *Educational Measurement in the Elementary Grades*, pp 137-50. Yonkers-on-Hudson, N Y World Book Co, 1930
- Mort, Paul R, and Gates, Arthur I, *The Acceptable Uses of Achievement Tests*, Chapter VII New York Bureau of Publications, Teachers College, Columbia University, 1932
- Morton, Robert L, *Teaching Arithmetic in the Intermediate Grades*. New York Silver, Burdett and Co, 1927
- Morton, Robert L, *Teaching Arithmetic in the Primary Grades*. New York Silver, Burdett and Co, 1927.
- Myers, Garry C, *The Prevention and Correction of Errors in Arithmetic*. Chicago Plymouth Press, 1925
- Nelson, M J, *Tests and Measurements in Elementary Education*, pp 120-33. New York The Cordon Co., 1939.

- Randall, Joseph H, "Corrective Arithmetic in Junior High School." *Educational Method*, 16 182-85, January 1937
- Smith, Henry L., and Wright, Wendell W, *Tests and Measurements*, Chapter V. New York Silver, Burdett and Co, 1928
- The Teaching of Arithmetic*. Tenth Yearbook of the National Council of Teachers of Mathematics. New York Bureau of Publications, Teachers College, Columbia University, 1935.
- Thorndike, Edward L, *The Psychology of Arithmetic*. New York. The Macmillan Co, 1922
- Tiegs, Ernest W, *The Management of Learning in the Elementary Schools*, Chapter IX. New York Longmans, Green and Co, 1937.
- Tiegs, Ernest W., *Tests and Measurements in the Improvement of Learning*, pp. 125-32, 167-71. Boston Houghton Mifflin Co, 1939.
- Washburne, Carleton, "One Reason Children Fail in Arithmetic" *Progressive Education*, 9 215-23, March 1932
- Webb, L. W., and Shotwell, Anna Markt, *Testing in the Elementary School*, Chapter X New York Farrar and Rinehart, Inc, 1939.
- Williams, Claude L, and Whittaker, R L, "Diagnosis of Arithmetic Difficulties." *Elementary School Journal*, 37 592-600, April 1937.
- Wilson, G. M., "Arithmetic" *Encyclopedia of Educational Research*, pp. 42-58. New York The Macmillan Co, 1941
- Wilson, Guy M., and Hoke, Kremer J, *How to Measure* (Revised and Enlarged Edition), Chapter IV. New York The Macmillan Co., 1929.
- Wilson, Guy M., Stone, Mildred B., and Dalrymple, Charles O, *Teaching the New Arithmetic*. New York McGraw-Hill Book Co., Inc., 1939.

CHAPTER XV

MEASUREMENT AND REMEDIATION IN THE RECEPTIVE LANGUAGE ARTS

The purpose of this chapter is to summarize the following important points involved in the identification and correction of instructional difficulties in the receptive language arts

- a.* Educational and social importance of reading.
- b.* Major objectives of reading instruction.
- c.* Problems of diagnosis in reading.
- d.* Testing reading readiness.
- e.* Testing and remediation in oral reading
- f.* Testing and remediation in work-type reading.
- g.* Illustrations of types of remedial material in reading.

The receptive language arts, which consist of reading and the study skills, are here distinguished from the expressive language arts, which include language, grammar, spelling, and handwriting. The receptive or assimilative language arts are dealt with in this chapter, while the following chapter presents the expressive or outgoing forms of the language arts. Treatments in both chapters are confined to the English language, because of the fact that the foreign languages are seldom taught in American elementary schools.

Educational and Social Importance of Reading Ability. The solution of most classroom problems in the modern school requires the skillful use of books as sources of information. When considered from this point of view, reading is something more than merely the rapid comprehension of printed symbols and the memory and organization of the materials read. It is also the ability to utilize books and libraries as efficient sources of information. This tendency to treat reading as a most important tool of learning has resulted in establishing a very close relation between reading and practically every other school activity. As a means of gaining information and pleasure it is essential in every content subject, such as history, geography, science, literature, and arithmetic.

In recent years there has also developed a keener appreciation of the importance of intelligent reading in society at large. Gray¹ states that reading is an indispensable means of "familiarizing adults with current events, with significant social issues, with community and national problems, and with American institutions, ideals, and aspirations. It is also essential in attaining vocational efficiency, in broadening one's range of information, and in seeking pleasure and profit during leisure hours." It is true that "of the making of books there is no end." The mass of printed matter which the average adult must read and evaluate, even within the limits of his own fields of interest, is stupendous. This situation makes the development of a high degree of reading skill in our schools all the more essential.

The necessity for a high level of reading ability on the part of all children and adults is more readily realized when it is recognized that a majority of the vast bulk of facts they are supposed to master are obtained from books, or at least as a result of reading. The real significance of the matter is seen in the fact that there is overwhelming evidence of a generally low level of reading ability on the part of these individuals who must use it so consistently.

I. IDENTIFICATION OF MAJOR READING ABILITIES

Major Objectives in Reading Instruction. The most common classification of reading objectives in the past has been to place them under the headings of oral and silent reading abilities. Recently it has seemed advisable further to cross-section these two types of reading skills and evaluate them in terms of their use in typical life situations. The life situations under which reading is usually done may be conveniently grouped into two types, depending on the attitude of the reader. These are: (1) reading for pleasure, or reading of the leisure type, and (2) reading for information, or reading of the work-level type. It must be clear that no hard and fast lines can be drawn between these two levels of reading activity, since an individual may shift from

¹W. S. Gray, "Importance of Intelligent Silent Reading" *Elementary School Journal*, 24 348-56, January 1924.

one type to the other without realizing the change. However, the distinction between the type of situations under which reading ordinarily takes place affords a convenient basis for the analysis and location of specific reading skills around which to develop diagnostic and remedial materials.

Outline of Basic Reading Skills. The outline of the field of reading given here is adapted from an outline of reading skills by Horn and McBroom,² with supplementation from a functional analysis by Yoakam³ and from an analysis of reading activities of high school pupils by Gray.⁴

A FUNCTIONAL ANALYSIS OF READING

A. Reading of Leisure Types

I. Oral Reading

1. Entertainment of others
2. Self-entertainment
3. Appreciation of beauty of expression
4. Practice on expression

II. Silent Reading

1. Self-improvement
2. Self-entertainment
3. Vicarious experience
4. Appreciation
5. Practice

B. Work-Types of Reading (major portion of reading time)

I. Oral Reading

1. Entertainment of others
2. Instruction of others
3. Acquisition of new modes of expression
4. Extension of vocabulary

II. Silent Reading (major aspects of work-types of reading)

1. Examples of specific situations in which one reads silently
 - a. To secure the facts on which judgments may be based
 - b. To secure information on the essential conditions of a problem to be solved

² Ernest Horn and Maude McBroom, *A Survey of the Course of Study in Reading* University of Iowa Extension Bulletin No 99. University of Iowa, Iowa City, February 1924.

³ G. A. Yoakam, *Reading and Study*, pp 54-56 The Macmillan Co., New York, 1928

⁴ W. S. Gray, "The Relation between Study and Reading" *Proceedings of the National Education Association*, 57 580-86. National Education Association, Washington, D. C., 1919

- c* To verify a fact or an opinion
- d* To secure a basis for action
- e*. To secure general information in desirable fields
- f*. To acquire information for specific use
- g*. To evaluate material as to its applicability to a specific case
- h*. To secure a basis for the formulation of an opinion from data or statements
- i*. To analyze the essential elements of an argument
- j*. To enlarge one's vocabulary
- k*. To acquire more effective modes of expression
- l* To master technical and contextual meanings of words
- m*. To acquire a basis for a critical questioning attitude on a controversial issue
- n*. To provide a basis for the understanding of a specific situation
- o*. To satisfy the desire to discover new problems and answer them

III. Essential Skills Involved in Work-Study Types of Reading, and Suggested Methods of Acquiring Them

- | | |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>1. Skills and Abilities</p> <ul style="list-style-type: none"> <i>a</i>. Skill in recognizing new words. <i>b</i>. Ability to locate material quickly <ul style="list-style-type: none"> (1) Knowledge of and ability to use an index. (2) Ability to use a table of contents (3) Ability to use the dictionary. | <p>2 Methods</p> <ul style="list-style-type: none"> <i>a</i>. Methods suggested in manuals of standard method readers. <i>b</i>. — <ul style="list-style-type: none"> (1) Children learn alphabet, find words in alphabetical arrangement, find words in an index, find answers to questions by use of index, make an index for a book which has none. (2) Assign lessons by title, find lessons in table of contents, find authors in table of contents. (3) Handled in detail in Rice, O S, <i>Lessons on the Use of Books and Libraries</i>, Rand, Mc- |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Nally and Com-
pany.

- | | |
|---------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (4) Ability to use library card files. | (4) Same as (b, 3). |
| (5) Ability to use reference material. | (5) Same as (b, 3). |
| (6) Ability to use keys, tables, graphs, etc. | (6) Practice in interpreting maps, tables, etc., in geographies, etc., make graphs or tables to illustrate problems |
| (7) Ability to skim. | (7) Certain topic given, children skim to find material about it, to find answer to a question, a sentence which proves point, etc. |
| c. Ability to comprehend quickly what is read. | c. Use flash cards with words, phrases, sentences, requiring action responses, find answers to questions, answer true and false statements, follow directions, completion sentences, labeling objects, matching words and objects, or words and pictures |
| (1) Rhythmic and rapid eye movements. | |
| (2) Absence of lip reading | |
| (3) Knowledge of meaning. | |
| d. Ability to select and evaluate material needed | d. Same as (c), also practice noticing date of publications, name of author, choose from several statements one which answers question best. |
| e. Ability to organize what is read | e. Practice giving subjects to a paragraph, choose paragraph which answers questions, find everything which bears on a topic, put related words together, give a summary, outline a lesson or page. |
| (1) To summarize. | |
| (2) To assign topics to proper order or place | |
| (3) To discover related material | |
| (4) To outline. | |

- | | |
|-----------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| <i>f.</i> Remembrance of material read. | <i>f.</i> Analysis of how to memorize quickly, practice in it, points in an article memorized; reports made without notes |
| <i>g.</i> Knowledge of sources. | <i>g.</i> Practice using common sources dictionary, encyclopedias, yearbooks, magazines, readers' guides, etc., make lists of sources. |
| <i>h.</i> Attitude of attacking reading with vigor. | <i>h.</i> Hold every child responsible for jobs assigned, tie up reading with project work, show each child his needs through testing. |
| <i>i.</i> Attitude of proper care of books. | <i>i.</i> Lessons on how books are bound, what ruins bindings, how to keep books clean, how to make them endure; who pays for the books, etc. |

The foregoing outline of the major objectives of reading instruction affords a useful basis for the evaluation of present instructional emphasis as well as a valuable list of criteria for the validation of diagnostic and remedial devices in reading.

II. GENERAL ANALYSIS AND DIAGNOSIS OF READING DISABILITIES

Typical Defects in Reading. The solution of the problem of the effective initial teaching of reading as well as the development of satisfactory remedial materials in reading is dependent to a large degree upon the accurate identification of the specific causes of reading failure. Not only is it necessary to discover the child who in his later school experience is almost certain to encounter reading difficulties, but these reading difficulties must be identified much more

definitely and accurately than has been the case in the past. Harris lists and discusses at length⁵ the following causes of reading difficulties: (1) low intelligence, (2) visual defects, (3) auditory defects, (4) other physical conditions—defects of muscular coordination and speech, illness, and glandular disturbances, (5) lack of hemispherical dominance, (6) emotional factors, and (7) inadequate home environment. He points out, however, that “in the majority of cases one cannot single out a particular handicap as the sole factor that is responsible for the child’s disability in reading. More often than not, investigation will show the presence of several handicaps, any one of which may have interfered with progress.”⁶

Oral vs. Silent Reading. An examination of the major aspects of remedial work in reading indicates that there are two angles from which it may be considered. In the first place, remedial instruction may be begun in the oral reading field. Judd, Gray, and others have defended this point of attack on the problem on the ground that it enables the teacher to start with the child on a level at which he already has some mastery, that is, the oral language level. Others believe that on account of the large proportion of reading time spent in the work-type of silent reading this field should receive the special emphasis. There is merit on both sides of the question, undoubtedly. It is true that the child does come to school with a fairly adequate oral vocabulary which in a great many ways affords the natural approach to reading. On the other hand, it is also true that such an approach tends to place too large an emphasis on the pronunciation of words and too little upon their meaning when encountered in silent reading situations. The transfer from the emphasis on oral reading (pronunciation of words) to silent reading (comprehension of meaning of words, sentences, and paragraphs) must be made at some point in the child’s experience. Accordingly, a great many teachers hold that the place to start the emphasis on silent reading is at the beginning. Some foundation for this belief is seen in the

⁵ Albert J. Harris, *How to Increase Reading Ability*, Chapter VI. Longmans, Green and Co., New York, 1940

⁶ *Ibid.*, p. 167

results obtained by many teachers who place the emphasis on the development of silent reading skills at the outset.

III. DETERMINATION OF READING READINESS

Factors in Reading Readiness. Reading readiness is dependent upon a large number of characteristics. Harris lists the following as among the most important: (1) intelligence, (2) visual perception, (3) auditory perception, (4) language development, (5) background of experience, and (6) social behavior.⁷ It is unsafe to assume that a child who enters school at the age of six is ready for reading. Some children have already learned to read, and are mentally much more mature than the average child of six, while others may have no more mental maturity than the average child of four, and thus encounter difficulty in learning to read.

Betts,⁸ reporting on the basis of results in a reading clinic, says:

It is imperative that a teacher should not drive a child into reading until she has made an attempt to analyze or define the problem. Our records show that almost 90 percent of the severe reading cases should have medical attention before receiving pedagogical help. In such instances, tutoring aggravates the problem and many times an apparent gain in reading achievement is due to maturation rather than to the pedagogical methods used.

Such cases depend for their effective remediation upon the services of persons other than the teacher, obviously, but they are no less a classroom concern for that reason.

Reading Readiness Tests. Reading readiness tests can best be classified as tests of specific intelligence, for their purpose is to measure the mental ability factors essential to success in reading. These factors are measured by tests which make use of visual and auditory abilities which are basic to reading. Reading readiness tests employ several testing devices. Among them are: (1) distinguishing pictured objects which are named by the examiner, (2) matching one word of a group with its counterpart, which appears as a

⁷ Harris, op cit, p 48

⁸ Emmett A. Betts, "Teacher Analysis of Reading Disabilities" *Elementary English Review*, 11 99-102, April 1934.

visual stimulus, (3) recognizing word similarities or differences, (4) recognition of rhyming words, and (5) reading numbers and letters. Only the last of these can perhaps be called a reading skill, although all of the types measure abilities upon which later reading abilities depend. The major purpose of reading readiness tests is to locate those children who are not yet ready to start reading but for whom that activity should be delayed until the child's mental maturity and experience are adequate for such an undertaking.

The accompanying illustrations from the Stone-Grover *Classification Test for Beginners in Reading* are representative of testing methods for reading readiness. Instructions are given orally by the examiner for all such tests. This test serves a classification and also a readiness testing function.


The *Gates Reading Readiness Test*, issued recently to accompany some of his reading tests which will be discussed in the following pages, consists of five parts, the last one of which must be administered to one child at a time. The five parts of the Gates test are: (1) picture directions, (2) word matching, (3) word-card matching, (4) rhyming, and (5) letters and numbers.

EXCERPTS FROM STONE-GROVER CLASSIFICATION TEST FOR
BEGINNERS IN READING⁹

TEST I

The child looks at the word under the picture, or in the "box" if there is no picture, finds the same word among the other words and draws a line under it. The score is one point for each correct response. Time: 7 minutes.

PREPARATORY EXERCISES

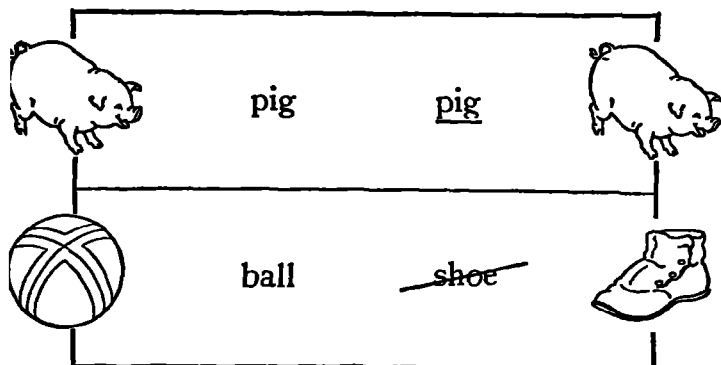
 <p>run</p>	jump	<u>run</u>
	cat	she

⁹ Clarence R. Stone and C. C. Grover, *Classification Test for Beginners in Reading*. Published by Webster Publishing Co., 1933.

TEST II

The child looks at the two words in the "box" and decides whether the second one is the same as the first or different. If the second word is the same as the first, he underlines the second. If the second is different from the first, he draws a line through the second, as shown below. The score is one point for each correct response. Time, 7 minutes.

PREPARATORY EXERCISES



Other reading readiness tests embody in general the same measurement devices illustrated and mentioned above, so no additional illustrations are given here. The *Lee-Clark Reading Readiness Test* requires pupils to match similar letters, to recognize the one letter which is unlike the other three of a group, and to cross out the extra letter occurring in one of two words otherwise alike. The *Monroe Reading Aptitude Tests* provide five types of tests: (1) visual, for measuring memory for position of visual forms, control of eye movements, and drawing from memory, (2) auditory, for measuring ability to detect correct pronunciations, to distinguish between words which sound somewhat alike, and to reproduce a story from memory, (3) language, for measuring extent and richness of vocabulary, (4) articulation, for measuring correctness of articulation and speed in repeating words, and (5) laterality, for measuring hand, eye, and foot preferences.

The *Metropolitan Readiness Tests*, the *Van Wageningen*

Reading Readiness Tests, the *Betts Ready to Read Tests*, and the *Durrell-Sullivan Reading Capacity Tests* are other tests which serve the double function of providing an indication of whether or not pupils are ready for reading instruction and of providing for those who are ready for instruction a basis for their classification or remediation.

IV. ANALYSIS AND DIAGNOSIS IN ORAL READING

Oral Reading Paragraphs. The *Gray Standardized Oral Reading Paragraphs* probably represent the most highly developed tests of oral reading ability. While they were not designed specifically for diagnostic purposes, the fact that they are used as individual tests and not as group tests makes them rather useful diagnostic instruments for the teacher of the lower elementary school grades. They are standardized for use in the first to the eighth grades, but the shift in the emphasis from oral to silent reading in the upper grades decreases the significance of their use in the higher grades. Paragraphs 1, 6, and 12 of the test are reproduced here as an illustration of the character of the material.

EXCERPTS FROM GRAY ORAL READING PARAGRAPHS¹⁰

I

A boy had a dog
The dog ran into the woods.
The boy ran after the dog
He wanted the dog to go home.
But the dog would not go home.
The little boy said,
"I cannot go home without my dog."
Then the boy began to cry.

6

The part of farming enjoyed most by a boy is the making of maple sugar. It is better than blackberrying and almost as good as fishing. One

¹⁰ William S. Gray, *Standardized Oral Reading Paragraphs*. Published by Public School Publishing Co.

There are 12 of these paragraphs arranged in increasing order of difficulty. The tests are administered individually by having the pupils read the paragraphs under time control until a certain number of errors per paragraph are made. Rate and quality of reading are then combined by means of a table of values provided for the purpose.

reason why a boy likes this work is that someone else does most of it. It is a sort of work in which he can appear to be very industrious and yet do but little.

12

The hypotheses concerning physical phenomena formulated by the early philosophers proved to be inconsistent and in general not universally applicable. Before relatively accurate principles could be established, physicists, mathematicians, and statisticians had to combine forces and work arduously.

Oral Reading Check Tests. The plan of recording the number and the kinds of errors made by the pupil in reading the *Gray Standardized Oral Reading Paragraphs* permits a type of diagnostic analysis of oral reading abilities. Much more concise information of this kind is made available, however, through the use of *Gray's Oral Reading Check Tests*.

As in the oral reading paragraphs, these check tests are to be given individually. The errors made by the pupil are recorded by the teacher on a separate test sheet showing the types of errors made by the pupil which appear most frequently in oral reading. The following illustration may make clear the character of the errors and the method of recording them:¹¹

The sun ^{many} pierced into my large windows. It was the opening of October, and the ^{clear} sky was of a dazzling blue. I looked out of my window and down the street. The white house of the long, straight street were almost painful to the eyes. The clear atmosphere allowed full play to the sun's brightness.

If a word is wholly mispronounced, underline it as in the case of "atmosphere." If a portion of a word is mispronounced, mark appropriately as indicated above "pierced" pronounced in two syllables, sounding long *a* in "dazzling," omitting the *s* in "houses" or the *al* from "almost," or the *r* in "straight." Omitted words are marked as in the case of "of" and "and"; substitutions as in the case of "many" for "my", insertions as in the case of "clear", and repetitions as in the case of "to the sun's." Two or more words should be repeated to count as a repetition.

¹¹ Ibid.

GRAY ORAL READING CHECK TEST, INDIVIDUAL RECORD SHEET¹²INDIVIDUAL RECORD SHEET
PROGRESSIVE ANALYSIS OF ERRORS IN ORAL READING

Pupil's Name _____ Age _____ Grade _____

Types of Errors	No 1	Daily	No 2	Daily	No 3	Daily	No 4	Daily	No 5	Daily
I. INDIVIDUAL WORDS										
1 Non recognition						- -				
2 Gross mispronunciation	-				-					- -
3 Partial mispronunciation										
a. Monosyllabic Words										
1 Consonant										
2 Vowel				-		-				-
3 Consonant blends	-								-	
4 Vowel digraph						-		-		
5 Pronounce silent letters								-		
6 Insert letters						-		-		
7 Pronounce backwards	- -									
8 Rearrange letters										
b. Polysyllabic Words										
1 Accent										
2 Syllabication										
3 Omit syllable										
4 Insert syllable										
5 Rearrange letters of syllables	-	-						-		-
6 Incorrect pronunciation of a syllable								-		-
4 Fnuunciation								-		-
5 Substitutions								-		-
6 Insertions								-		- -
7 Omissions								-		
8 Other types of error {	-	-		-		-		-		-
	-	- -								
II GROUPS OF WORDS										
1 Change order	-	-				- -		-		
2 Add words to complete meaning according to fancy										
3 Omit one or more lines										
4 Insert two or more words										
5 Omit two or more words							-		-	
6 Substitute two or more words										
7 Repeat two or more words										
8 Other types of error {	-	-				-				
	-	- -								
Pupil's test record { Rate Errors	-							-		
Standard Scores for the Grade { Rate Errors			- - -						-	
Date of Each Test	- -		- -		-			- -		

¹² William S Gray, *Standardized Oral Reading Check Tests*, Individual Record Sheet Published by Public School Publishing Co.

The individual record sheet which accompanies Gray's *Oral Reading Check Tests* is useful in two important ways. It places before the teacher a carefully classified list of common errors in oral reading, and it provides space for the recording of successive repetitions of the test so that progress may be measured.

The analysis of the individual pupil's record gives a very concise picture of his oral reading difficulties. It will be noted that in these oral reading exercises no attention is paid to the degree of comprehension with which the material is read. The measurement of comprehension lies somewhat beyond the purpose of this test. Here the purpose is the determination of the efficiency with which words are recognized and pronounced in context, with little or no concern for the comprehension of the materials.



V. ANALYSIS AND DIAGNOSIS IN SILENT READING



Measurement of Work-Study Type of Reading. The emphasis given to the work types of reading in the list of skills given on pages 328 to 331 indicates something of the importance of this type of reading in relation to the total reading field. Some pupils fail (in arithmetic, for example), not entirely because of ignorance of the basic facts, or lack of mental ability to understand the explanations, but rather on account of sheer inability to read. In fact, one of the best ways to improve work in many other school subjects is to make a drive on the work-type of reading ability. A recognition of this has caused makers of tests in reading to turn their attention in this direction in recent years. A number of excellent reading tests which provide useful analytical information concerning a number of work-study skills are available.

The *Gates Silent Reading Tests* are prepared in two series and two forms for use in the primary grades and in the intermediate grades. The primary tests are available in Types 1, 2, and 3, while four types, Types A, B, C, and D, of the intermediate test are available. A single exercise from each of these types is given as an example of the content of these tests.







EXCERPTS FROM GATES PRIMARY READING TESTS¹⁸

TYPE 1 WORD RECOGNITION














	did	egg
	dog	two
	be	bed
	bag	she

	may	make
	come	milk
	horse	play
	hose	house

TYPE 2. SENTENCE READING

This is a cat. I			
This is a book. II			
This is a cup. III			

TYPE 3 READING OF DIRECTIONS

   <p>1 Put an X on the ball</p>	   <p>3. Draw a line under the little book</p>
     <p>2 Put an X on the milk bottle</p>	  <p>4. Draw a line from the pig to the tree</p>

¹⁸ Arthur I. Gates, *Gates Primary Reading Tests*. Published by Bureau of Publications, Teachers College, Columbia University, 1926 and 1931.

EXCERPTS FROM GATES SILENT READING TESTS¹⁴

TYPE A. READING TO APPRECIATE THE GENERAL SIGNIFICANCE

Once upon a time a young fairy went down to the river to swim. She jumped in with a splash. She put out her hands and tried hard to swim. Something seemed to be dragging her down. Oh, it was her wings! She had forgotten to take them off. Fairy wings become heavy when they are wet. She cried for help as loudly as she could.

Draw a line under the word which tells how the fairy felt

cross angry weary afraid joyful

TYPE B. READING TO PREDICT THE OUTCOME OF GIVEN EVENTS

The grocery man had a black cat. He loved his cat very much. One day a lady brought a big bulldog into the store. The grocer's cat raised his back and said "Meow! Psst!" to the bulldog. Of course, the dog did not like that, so he growled loudly. Before the grocery man or the lady knew what was happening, the bulldog had sprung upon the cat.

They let the fight go on
The cat slept on
The lady took her bird away
The grocery man saved his cat

TYPE C. READING TO UNDERSTAND PRECISE DIRECTIONS



¹⁴ Arthur I. Gates, *Gates Silent Reading Tests*, Grades 3 to 8. Published by Bureau of Publications, Teachers College, Columbia University, 1926.

- | | |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>3. The third grade boys have a baseball team. It is made up of the best nine players in that grade. They have played seven games this spring and have won five. They hope to win all the others. Draw a line under the number that tells how many games they have played.</p> | <p>7. This miner is getting ready to go down into a gold mine. He has on a miner's hat with a torch on the front of it. He lights it just before he goes into the dark tunnels. Make a cross where the miner will carry his light.</p> |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

TYPE D. READING TO NOTE DETAILS

Next morning she awoke and found herself in a beautiful room. The walls were covered with silken curtains. There were two mirrors made of pure silver. The bed was made of ivory. The coverings were made of silk and velvet. By her bed lay a dress and a pair of slippers. The dress was made of silk. The slippers were covered with diamonds.

Where did the girl find herself?
 barn room garden store
 What were the mirrors made of?
 silver gold pearl silk
 What were on the slippers?
 rubies pearls opals diamonds

The *Iowa Silent Reading Tests, New Edition, Elementary*, are among the more recent and comprehensive tests designed to provide a detailed and analytical measure of silent reading abilities. These new quick-scoring tests go beyond the general survey of two or three phases of silent reading ability. They cover a wide range of skills essential to effective reading of the work-study type. Naturally they do not succeed in measuring all of the major objectives of reading as outlined in the opening pages of this chapter. The specific skills which are measured by the elementary tests are given on page 56 of this book.

These tests sample into numerous reading skills and into

different subject-matter fields. For example, Test 1 consists of two articles, one dealing with science and the other with history content. Rate of reading is measured by these two samples under definite comprehension requirements. The pupil is told to read the articles as rapidly as possible and yet be able to answer certain comprehension questions based on the content. Test 2, *Directed Reading*, utilizes the same two articles of science and history content for an intensive check on the pupil's ability to comprehend certain questions and to locate their answers in the articles. Test 3, *Word Meaning*, contains two groups of exercises. The first group samples into general vocabulary, and the second into subject-matter vocabulary in mathematics, science, and social science arranged in cyclic order. Test 4 measures three phases of paragraph comprehension.

One of the more valuable reading skills measured by these tests is the ability to use the index for the purpose of locating information. This test, which is Part B of Test 6, is reproduced here in order to give a suggestion of how this ability may be measured. It should also give to many teachers hints for the development of similar material for instructional purposes.

The *Diagnostic Examination of Silent Reading Abilities* is another highly diagnostic instrument. It measures ten aspects of silent reading ability and furnishes a reading index which is a specialized type of derived score. Illustrations from the parts measuring rate of comprehension and various types of reading comprehension are given on page 346. The following list of the part scores obtained for the test indicates its highly diagnostic nature and its coverage of specific types of high level work-study skills: (1) rate of comprehension, (2) perception of relations, (3) vocabulary in context, (4) vocabulary—words in isolation, (5) range of general information, (6) central thought, (7) clearly stated details, (8) interpretation, (9) integration of dispersed ideas, and (10) drawing inferences. From the various items of Part III, for which one paragraph and the accompanying test items are illustrated, scores of the types listed in (6) to (10) above are obtained by means of the scoring method used.

EXCERPT FROM IOWA SILENT READING TEST,
NEW EDITION, ELEMENTARY 15

Iowa Silent Reading: New Ed. Elem. A-B

TEST 6. PART B: USE OF AN INDEX

DIRECTIONS The answers to the questions in Column 2 are found in the index below. First read the question and then find the desired answer by looking under the proper topic in the index. Then locate your answer among the possible answers given with the question and fill in the answer space in the margin which is numbered the same.

Study the samples carefully before you try to answer the questions

INDEX

Canada 45-53, coal, 244; dairying, 157; flax, 153; forests, 97; fur farms, 176-177; industrial regions, 263; map, 47; trapping 176; wheat, 95-98

China 22-24, deserts, 216; farming 125; 120-131, lack of dairy cattle, 130, lack of manufacturing, 262; plains, 129; population, Fig. 24; rice, 131; silk, 131; 144, troubled condition 124

Cotton 147-152, Australia, 151, bales, 149, bolls, 149, Egypt, 150, Korea, 150, mills, 149-150, Russia in Asia, 152, South Africa, 154 Sudan, 156, United States production, Fig. 42, page 148

Dairying 156, in Holland, 158

Geology 20-24, defined, 21, important divisions, 21, fields, 24

Germany 83-85, 235, cities, 82, dairying, 158, farming, 83, fisheries, 203, manufacturing, 267, potatoes, 84, rainfall, 82, sugar beets, 82

Grapefruit California, 135, Florida, 133-134, Texas, 136

Herdin Persu, 219, rancher in Alaska, Fig. 25, page 102, reader in Lapland, 102, Russia in Asia, 219, semi deserts, 213-214. See also Grazing

SAMPLES

- A. What page discusses lumbering in Oregon?
1 138 2 173 3 92 4 98 5 186
 - B. What page gives information about fisheries in Germany?
1 166 2 235 3 83 4 205 5 82
-
1. Next to what page can you find a map of Japan?
1 127 2 142 3 206 4 269 5 282 1
 2. What page tells about dairy products in Holland?
1 127 2 131 3 156 4 157 5 168 2
 3. Does the index tell where to find the number of miles of railroad in the United States?
1 Yes 2 No 3
 4. What is the number of the figure which tells about cotton production in the United States?
1 24 2 39 3 42 4 45 5 196 4
 - 5 Where is there a reference to grapefruit in Texas?
1 trade 2 grapefruit 3 Texas 4 fruit 5
 - 6 On what page can a definition of geology be found?
1 21 2 22 3 23 4 24 5 25 6
 7. What page tells about the wheat crop in Oregon?
1 95 2 98 3 99 4 138 5 206 7
 8. Where (on what page) will you find a map of Canada?
1 22 2 45 3 47 4 97 5 176 8

India: 22-27; animals, 93; barley, 90; cotton, growth of, 112; crowding, 90; dairying, 161; fairs, 91; farmhouses, 92; farming, 90; governed by England, 90-91; rainfall, 90; rice, 90; seasons, 90; silk, 141	1	2	3	4	5	6	7	8	9	10	11	12
Japan 267-269; charms, 132; farming, 127-128; fisheries, 206; forests, 127; map opposite page 269; poverty, 128; rice, tea, 142; temperate belt, 127	1	2	3	4	5	6	7	8	9	10	11	12
Oregon apples, 138; automobiles, 99; horses, 99; irrigation, 138; lumber, 173; salmon, 205, 207; wheat, 98	1	2	3	4	5	6	7	8	9	10	11	12
Railroads 41-45; Alaska, 185; deserts, 211-212; east of Caspan Sea, 104; Lapland, 194; number of miles in United States, 45; Moscow, 196; spur track, 39 See also Transportation.	1	2	3	4	5	6	7	8	9	10	11	12
Scandinavian Peninsula - 195-196; mountains, 193 See also Norway	1	2	3	4	5	6	7	8	9	10	11	12
Science defined, 18; American Men of, 22	1	2	3	4	5	6	7	8	9	10	11	12
South Africa 94, 105-110; cotton, 151; oranges, 138; ranching, 163; sugar, 155	1	2	3	4	5	6	7	8	9	10	11	12
Texas map of, 75; admission to Union, 78	1	2	3	4	5	6	7	8	9	10	11	12
Trade 282-288; advantages for, 283-284; Arctic Ocean, 196; Eskimos with white people, 188; Hawaiian sugar, 115	1	2	3	4	5	6	7	8	9	10	11	12
Venice 23, 280-282; Grand Canal, 280; manufacturing, 282	1	2	3	4	5	6	7	8	9	10	11	12
Look at Sample A In the index under "Oregon" you will find the word "lumber" and the page reference, 173 173 is second among the answers given with the question, so the second answer space has been filled in	1	2	3	4	5	6	7	8	9	10	11	12
Look at Sample B See if you can find the page reference in the index The correct answer space is marked	1	2	3	4	5	6	7	8	9	10	11	12
Answer the remaining exercises the same way.	1	2	3	4	5	6	7	8	9	10	11	12
9. Where is English government in India discussed? 1 23-27 2 83-85 3 90-91 4 92-93 5 133-134	1	2	3	4	5	6	7	8	9	10	11	12
10. Does the index tell where to find something on the schools of India? 1 Yes 2 No	1	2	3	4	5	6	7	8	9	10	11	12
11. On what page will you find a figure showing how many people live in China? 1 24 2 124 3 125 4 126 5 129 11	1	2	3	4	5	6	7	8	9	10	11	12
12. What topic gives another reference to herding? 1 cattle 2 dairying 3 farming 4 grazing 12	1	2	3	4	5	6	7	8	9	10	11	12
13. On what page is ranching in South Africa discussed? 1 106 2 138 3 151 4 155 5 163 13	1	2	3	4	5	6	7	8	9	10	11	12
14. On what page can you find a figure about herding reundeer in Alaska? 1 192 2 194 3 213 4 214 5 219 14	1	2	3	4	5	6	7	8	9	10	11	12
15. Where (on what page) is a map of Texas given? 1 47 2 75 3 78 4 92 5 193 16	1	2	3	4	5	6	7	8	9	10	11	12
16. Where will you find out about cotton in Egypt? 1 147 2 148 3 149 4 150 5 154 16	1	2	3	4	5	6	7	8	9	10	11	12
17. Is there a reference in this index to British Men of Science? 1 Yes 2 No 17	1	2	3	4	5	6	7	8	9	10	11	12
18. On what page are the Industries of Canada referred to? 1 153 2 157 3 176 4 263 5 277 18	1	2	3	4	5	6	7	8	9	10	11	12

Do not turn this page until you are told to do so.

Number right

EXCERPTS FROM DIAGNOSTIC EXAMINATION OF SILENT READING
ABILITIES, INTERMEDIATE¹⁶

Part I

A. John's car came to a stop because there was no more gasoline in the tank. When he had to walk over a mile to get ~~water~~, it made him cross.

In the last half of this paragraph, the word WATER does not fit in with the meaning of the rest of the paragraph, so WATER is crossed out

B The carpenter asked Tom to go to the hardware store and get him a pound of nails. When Tom got back with the matches, the carpenter gave him a nickel.

C. We are planning to go on an all-day picnic tomorrow. We want to get started just as early in the afternoon as we can get away.

Part III

11. The telephone is made of many different things. The wires are made of gold, silver and platinum melted together. The receiver case is made of rubber. The part into which you speak, called the transmitter, contains aluminum, mica, nickel, coal, and a paper made from linen. Iron, copper, tin, and zinc are used on the inside of the receiver, the part you hold to your ear. Shellac is used in making the mouth-piece.

48. The paragraph is mainly about

41. metals 42. telephone wires 43. making the telephone

44. parts of the telephone 45. what telephones are made of 48

49. Which of these is used in making the receiver?

46. gold 47. tin 48. nickel 49. linen 50. mica - - 49

50. The part of the telephone through which you hear is called the

51. wire 52. mouth-piece 53. receiver case 54. receiver

55. transmitter - - - - - 50

Performance Tests in Silent Reading. The *Betts Ready to Read Tests* include not only reading readiness tests but also tests for the diagnosis of difficulties for pupils who do not read normally. The tests for oculomotor and perception habits require the use of a series of slides and the *Betts-Keystone Telebmocular*,¹⁷ a type of stereoscope which provides a scaled holder adjustable for various distances. The tests measure fusion, visual acuity, muscular balance, eye-coordination, depth perception, and astigmatism. Their pur-

¹⁶ M. J. Van Wagenen and August Dvorak, *Diagnostic Examination of Silent Reading Abilities*, Intermediate. Published by Educational Test Bureau, 1939.

¹⁷ Distributed by Keystone View Co., Meadville, Penn.

pose is not to diagnose visual defects as a basis for prescription but to locate pupils who should be referred to eye specialists for examination and remediation.

The *Ophthalmograph*¹⁸ is a binocular eye-movement camera used to obtain a simple and objective record of eye movements during the reading process. Information is provided on a film strip concerning the number of eye fixations, recognition span, regressive eye-movements, rhythm, reading speed, and coordination of the eyes. Charts are provided for use in easy determination of total reading time for a given number of words. This procedure measures the eye mechanics of reading, and should ordinarily be supplemented by a test of reading comprehension.

The *Metronoscope*¹⁹ is a device for exposing printed strips of reading matter at desired rates of speed and can be used either with individuals or small pupil groups for testing and drill purposes.

The *Durrell Analysis of Reading Difficulty* materials include a hand-operated tachistoscope, or device for exposing reading strips at desired rates, for use in determining word recognition and phrase comprehension. A test of oral reading measures phrase reading, voice, enunciation, expression, and general word skills, and is accompanied by questions to test comprehension.

VI. CORRECTIVE EXERCISES IN READING

Remedial Drills for Oral Reading Difficulties. Dearborn, Huey, Gray, Buswell, and many others, studying the problem of how to improve reading, have noted that there is a marked relationship between the rate and quality of children's reading and the control they have over their eye-movements in reading. The meaning of eye-movements may be readily understood by anyone who will take a position closely in front of and directly in the range of vision of a person engaged in reading. The observer will note that the reader's eyes do not move regularly and systematically forward as the reading progresses but that the movements

¹⁸ Distributed by the American Optical Co., Southbridge, Mass.

¹⁹ *Ibid*

are interspersed with pauses or fixation periods. It is during these pauses that the images of the words or groups of words are secured. Carefully conducted laboratory experiments reveal the fact that good readers make longer sweeps with the eyes, take in larger units of words, pause for much shorter periods, and rarely retrace material once covered. Gates²⁰ concludes that improper eye-movements are probably the evidences of other types of reading disability which can be treated specifically. With the removal of the causes lying back of ineffective eye-movements, the treatment of eye-movements as such becomes unnecessary. On the whole this seems to be the most hopeful way of looking at the problem, since a great many teachers are qualified to administer types of remedial treatments which may be applied in the classroom but only a few have the technique or equipment for training the pupil in more effective eye-movements. Gray himself recognizes this practical aspect of a problem of remedial instruction in reading and suggests a number of excellent exercises designed to overcome specific difficulties in reading, many of which are indicated by the ineffectual eye-movements of the pupil. One of these exercises is reproduced here to illustrate types of material adapted to remedying certain oral-reading difficulties.

EXERCISE TO INCREASE ACCURACY OF RECOGNITION²¹

1. Words which a pupil failed to recognize accurately while reading were used in sentences at the end of each period, in order that he might associate them with their meaning. The words which repeatedly caused difficulty were then typewritten on cards and used in quick-perception drills, by presenting them as rapidly as they were recognized. Such words as *again*, *want*, *been*, *does*, and *heard* were frequently emphasized. As soon as a pupil was able to recognize a word readily, drill on it was discontinued. New words were added to the list as difficulties were encountered.
2. Words which a pupil confused because of their similarity in form were emphasized in drill exercises. These words included such groups as *thought*, *though*, and *through*, *there* and *where*, *then* and *when*, *now*

²⁰ Arthur I. Gates, *The Improvement of Reading: A Program of Diagnostic and Remedial Methods* (Revised Edition), pp. 338-40. The Macmillan Co., New York, 1935.

²¹ W. S. Gray, *Remedial Cases in Reading: Their Diagnosis and Treatment*. Supplementary Educational Monograph No. 22. University of Chicago Press, Chicago, 1922.

and *how*, and *has*, *had*, and *have*. The words were used in sentences before they were presented in quick-perception drills. If unusual difficulties were encountered, words which were similar in form were presented together so that their differences could be studied.

Pupils who recognized isolated words accurately frequently made errors in recognizing the same words in phrases and sentences. In order to overcome this difficulty a word, such as *there*, was written on the board in several phrases or short sentences and the pupil was given opportunity to study them deliberately. As soon as he was able to recognize these phrases readily they were typewritten on cards and presented in quick-perception drills.

By practice in the use of such diagnostic reading devices as these and by training themselves in careful observation of their pupils, teachers can become adept in the detection of particular reading difficulties. Furthermore, they will soon find that with practice they can become proficient in the art of building drill exercises. The sample exercises selected from a great many suggested by Gray, Gates, and others will be found valuable guides in the preparation of such materials. By following the examples given here, the teacher can be practically certain that he is using reading drill material whose efficiency has been experimentally established.

Remedial Drills for Work-Study Types of Reading Difficulties. Possibly one reason for the rather marked instructional emphasis on the work-study type of reading to the exclusion of reading of the leisure types is that it is reasonable to expect a considerable carry-over of skills from work-type reading to the other type, because of the large number of common skills involved. For example, skill in recognition of word meaning which functions in work-study reading is probably similarly effective when the individual is reading solely for pleasure.

A few illustrations of specific types of remedial exercises suited for use in silent reading of the work-study types are presented in the following pages in the hope that they may serve as a guide to teachers interested in the development of material of this type. Only a few samples from each field can be furnished.

Word Recognition. Exercises designed to develop skill in the recognition of new word meanings appear in a great

many forms, as for example: (1) simple sentence completion, (2) agent-action, (3) action-agent, (4) action-effect; (5) effect-action, (6) identification, (7) opposites, (8) similars, (9) description, (10) phrasing; etc. Exercises of these types are provided in the Horn-Shields readers²²

Location of Information in Books. A significant factor in the child's use of reading for work-study periods is his ability to locate information in books. It seems reasonable to suppose that the following suggestions may prove helpful.²³

1. To develop in pupils an ability to use the index, children should (a) be taught the alphabet, (b) be drilled in arranging words in alphabetical order, (c) be drilled in finding answers to questions by use of the index, (d) be asked to prepare an index for books not provided with them
2. To develop the ability to use a table of contents, pupils should (a) be assigned lessons by topic or titles, (b) find the assigned lessons in the text by means of the table of contents, (c) find additional sources of information on the assignment in the library

Organization of Material. The ability to organize what is read is a necessary part of the equipment of everyone who expects to become a good student. Organization of reading materials calls for a superior type of judgment. The following suggestions will aid teachers in developing a variety of types of practice in organization:

1. Practice in deciding upon the main thought in the paragraph or topic.
2. Drill in outlining a study, an assignment, a reference reading, a poem, etc
3. Practice in analyzing the organization of selections.
4. Practice in restating the substance of a difficult passage to convey the same idea in simplified form
5. Practice in selecting the most appropriate title for a selection

Other Remedial Materials in Reading. The recent analytical work of Gates, Gray, and others opens up great possibilities for diagnostic and corrective work. No longer need diagnosis in reading be confined to such vague and general qualities as rate, comprehension of word meanings, etc. In fact, it now becomes quite apparent that many of

²² Ernest Horn and Grace Shields, *Progress in Reading*. Ginn and Co, Boston, 1940

²³ O S Rice, *Lessons on the Use of Books and Libraries* Rand, McNally and Co, New York, 1920

the so-called diagnostic tests in this field are not at all suited for the specific types of diagnosis required in the identification of reading disabilities. While considerable progress has been made in the last decade in the more exact analysis and identification of underlying causes of reading disability, the development of adequate initial instructional materials and corrective devices has not kept pace with the analytical work. The next decade is almost certain to see much progress along these lines.

Commercial materials designed for instructional testing purposes and remedial uses in reading are available in many different forms. At present there are almost countless drill books and work-books having for their purpose the development of silent reading skills of the work-study type. However, this material has mainly emphasized the problems of teaching beginners to read by some particular method rather than that of providing corrective treatment for some basic disability.

TOPICS FOR DISCUSSION

1. Why would you expect reading disability to be reflected in classroom achievement?
2. In what specific ways does modern life place a particular burden on the ability to read rapidly and well?
3. Prepare a comprehensive list of oral reading objectives
4. Check the list of Essential Skills Involved in Work-Study Types of Reading against the skills specified for measurement in the Iowa Silent Reading Tests
5. Summarize your position with respect to the relative placement of instructional emphasis on oral and silent reading.
6. Study the type of index test shown on pages 344 and 345 and prepare a set of suggestive drill exercises for the development of skill of this type
7. Classify the major reading defects according to type.
8. Are improper eye-movements in reading causes of reading deficiencies or merely evidences of such defects?
9. Amplify the suggestions given for the increase in the pupil's span of attention in reading
10. Is there any special reason to assume that remedial practices which are helpful in curing oral reading defects may not also be useful in correcting silent reading deficiencies?

SELECTED REFERENCES

- Barker, Vilda, "Informal Testing of the Use of Books and Libraries" *Elementary English Review*, 10 143ff, 174ff., 205-8, June, September, and October 1933
- Betts, Emmett A, *The Prevention and Correction of Reading Difficulties*. Evanston, Ill Row, Peterson and Co., 1936.
- Betts, Emmett A, "Prevention and Correction of Reading Disabilities." *Elementary English Review*, 12 25ff, February, 1935.
- Broom, M. E, *Educational Measurements in the Elementary School*, pp 182-203. New York McGraw-Hill Book Co, Inc, 1939
- Brueckner, Leo J, and Melby, Ernest O, *Diagnostic and Remedial Teaching*, Chapter VIII Boston. Houghton Mifflin Co, 1931
- Crider, Blake, "Diagnosing Special Disabilities in Reading" *Educational Method*, 15 308-10, March 1936
- Dolch, Edward W, "First Step in Remedial Reading." *Elementary School Journal*, 37 268-72, December 1936.
- Dolch, Edward W, "Mass Remedial Reading." *Educational Administration and Supervision*, 23 541-46, October 1937
- Durrell, Donald D, *Improvement of Basic Reading Abilities*. Yonkers-on-Hudson, N Y World Book Co., 1940
- Fitzgerald, James A, "Diagnostic and Remedial Program in Reading" *Educational Method*, 17 221-25, February 1938.
- Gates, Arthur I, "Diagnosis and Treatment of Extreme Cases of Reading Disability" *The Teaching of Reading A Second Report* Thirty-Sixth Yearbook of the National Society for the Study of Education, Part I, Chapter XIII, pp. 391-416. Bloomington, Ill Public School Publishing Co, 1937
- Gates, Arthur I, *The Improvement of Reading A Program of Diagnostic and Remedial Methods* (Revised Edition). New York The Macmillan Co, 1935
- Gates, Arthur I, "The Measurement and Evaluation of Achievement in Reading" *The Teaching of Reading A Second Report* Thirty-Sixth Yearbook of the National Society for the Study of Education, Part I, Chapter XII, pp 359-88. Bloomington, Ill Public School Publishing Co, 1937
- Gates, Arthur I, *A Reading Vocabulary for the Primary Grades* (Revised and Enlarged Edition) New York Bureau of Publications, Teachers College, Columbia University, 1935.
- Gates, Arthur I, et. al, *Methods of Determining Reading Readiness*. New York Bureau of Publications, Teachers College, Columbia University, 1939
- Gilliland, A R, Jordan, R H, and Freeman, Frank S, *Educational Measurements and the Class-Room Teacher* (Revised Edition), Chapter VII New York The Century Co, 1931
- Gray, William S, "Contributions of Research to Special Methods Reading" *The Scientific Movement in Education* Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chap-

MEASUREMENT IN RECEPTIVE LANGUAGE ARTS 353

- ter VII, pp. 99-106. Bloomington, Ill.: Public School Publishing Co., 1938
- Gray, William S, "Reading" *Encyclopedia of Educational Research*, pp. 891-926 New York The Macmillan Co., 1941
- Gray, William S (Chairman), *The Teaching of Reading A Second Report*. Thirty-Sixth Yearbook of the National Society for the Study of Education, Part I. Bloomington, Ill.: Public School Publishing Co., 1937.
- Harris, Albert J, *How to Increase Reading Ability*. New York Longmans, Green and Co., 1940.
- Harrison, M. Lucile, *Reading Readness*. Boston Houghton Mifflin Co., 1936
- Herrick, Virgil E., "Selecting the Child in Need of Special Reading Instruction." *Elementary School Journal*, 40 424-34, February 1940.
- Lee, J Murray; Clark, Willis W, and Lee, Dorris May, "Measuring Reading Readiness." *Elementary School Journal*, 34 656-66, May 1934.
- McKee, Paul, *Reading and Literature in the Elementary School*. Boston: Houghton Mifflin Co., 1934.
- Madsen, I N, *Educational Measurement in the Elementary Grades*, pp. 150-60. Yonkers-on-Hudson, N. Y World Book Co., 1930
- Monroe, Marion, "Diagnosis and Treatment of Reading Disabilities." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XII, pp. 201-28. Bloomington, Ill. Public School Publishing Co., 1935
- Monroe, Marion, "Diagnostic and Remedial Procedures in Reading." *Educational Record*, 19 105-13, Supplement No. 11, January 1938.
- Monroe, Marion, et al, *Remedial Reading* Boston Houghton Mifflin Co., 1937.
- Mort, Paul R, and Gates, Arthur I, *The Acceptable Uses of Achievement Tests*, Chapter V New York Bureau of Publications, Teachers College, Columbia University, 1932
- Nelson, M. J, *Tests and Measurements in Elementary Education*, pp. 109-19. New York The Cordon Co., 1939
- Pond, Frederick L, "Qualitative and Quantitative Appraisal of Reading Experiences" *Journal of Educational Research*, 33 241-52, December 1939.
- Sangren, Paul V, *Improvement of Reading through the Use of Tests*. Kalamazoo, Mich Western State Teachers College, 1932
- Sangren, Paul V, "Methods of Diagnosis in Reading" *Elementary English Review*, 7 105-7, April 1930
- Smith, Henry L, and Wright, Wendell W, *Tests and Measurements*, Chapter IX. New York Silver, Burdett and Co., 1928
- Smith, Nila B., *American Reading Instruction* New York. Silver, Burdett and Co., 1934.
- Thorndike, Edward L, *A Teacher's Word Book of the Twenty Thousand Words Found Most Frequently and Widely in General Reading*

- for Children and Young People*. New York: Bureau of Publications, Teachers College, Columbia University, 1931.
- Tiegs, Ernest W., *The Management of Learning in the Elementary Schools*, Chapter VI. New York: Longmans, Green and Co., 1937.
- Tiegs, Ernest W., *Tests and Measurements in the Improvement of Learning*, pp. 110-25, 159-65. Boston: Houghton Mifflin Co., 1939.
- Traxler, Arthur E., *The Use of Test Results in Diagnosis and Instruction in the Tool Subjects* (Revised). Educational Records Bulletin No. 18, pp. 19-25, 49-56. New York: Educational Records Bureau, January 1937.
- Webb, L. W., and Shotwell, Anna Markt, *Testing in the Elementary School*, Chapters VIII-IX. New York: Farrar and Rinehart, Inc., 1939.
- Wilson, Guy M., and Hoke, Kremer, J., *How to Measure* (Revised and Enlarged Edition), Chapter V. New York: The Macmillan Co., 1929.
- Witty, Paul A., "Approach to Better Reading: An Evaluation" *Educational Administration and Supervision*, 25: 81-92, February 1939.
- Witty, Paul A., and Kopel, David, "Evaluating Reading and Remedial Reading." *English Journal*, 26: 449-58, June 1937.
- Witty, Paul A., and Kopel, David, "Motivating Remedial Reading: The Interest Factor" *Educational Administration and Supervision*, 22: 1-19, January 1936.
- Witty, Paul A., and Kopel, David, "Preventing Reading Disability: The Reading Readiness Factor" *Educational Administration and Supervision*, 22: 401-18, September 1936.
- Wrightstone, J. Wayne, "Diagnosing Reading Skills and Abilities in the Elementary Schools." *Educational Method*, 16: 248-54, February 1937.

CHAPTER XVI

MEASUREMENT AND REMEDIATION IN THE EXPRESSIVE LANGUAGE ARTS

This chapter presents a summary of the following points concerning measurement at the elementary school level in the fields of expressive language arts—language, spelling, and handwriting:

- a.* Analysis of specific language skills.
- b.* Measurement of oral and written language skills.
- c.* Remedial instruction in language.
- d.* Construction of spelling tests.
- e.* Diagnosis and remediation in spelling.
- f.* Measuring quality and rate of handwriting.
- g.* Diagnosis and remediation in handwriting.

The expressive language arts, consisting of language, spelling, and handwriting, are discussed in this chapter. These, together with reading and the work-study skills discussed in the preceding chapter, round out the language arts taught in the elementary school.

I. IDENTIFICATION OF LANGUAGE ABILITIES

The Social Importance of Language. The importance of language in everyday life naturally gives it a place of prominence among the instructional problems faced by the classroom teacher. It is significant that language, which was one of the first subjects to be measured, is one of the slowest to respond to analysis, diagnosis, and remedial treatment. Possibly this is due in part to the formal methods of instruction in this subject followed by many teachers. It is more likely to be due, however, to the sheer complexity of the subject itself, and the many forms in which it expresses itself.

In this discussion, language skill is considered to mean facility in the use of the proper language habits and forms essential to effective intercommunication. Such a point of view makes reasonably clear the problems of the language

teacher. Language skills arise, as do other specific skills, through the proper exercise of the desired habits. It goes without saying that proper exercise is possible only when proper identification of the habits has taken place. One cannot indulge in exercises until he knows what to exercise. Hence the habits upon which language skill depends must be identified, and much carefully constructed instructional and drill material must be provided. This in itself will serve two useful purposes. First, the use of good language drill material insures that the pupil will have experience in making the correct response to selected language situations either with or without the assistance of such formal grammar instruction as may be applied. Second, the use of such material sets up in the pupil's mind an attitude toward language error. A definite consciousness toward language errors must be developed. When a person becomes sensitive to these, and an expression such as "he don't" causes a reaction similar to that created in a cat confronted by a strange dog, he is on the way to rapid improvement in his language habits.

Analysis of Language Skills. It must be evident from what has been presented in earlier chapters in this book that an accurate analysis of the underlying skills in language is necessary before any significant program of diagnostic and corrective work can be undertaken. In the past, certain general language abilities have been identified for measurement purposes, such as language usage, grammar, composition, etc. In more recent years, however, there has been an effort to reduce language in general to its more elementary or basic skills.

The task of clarifying the statements and purposes of language (English) instruction and of analyzing and identifying the basic language skills is not one which will be successfully accomplished by any one person. At the outset it must be recognized that there are many conditions under which language functions. There is undoubtedly a language of impression, or comprehension, as well as a language of expression. The former is the aspect of language which is given particular attention in reading instruction. The latter, the language of expression, is the phase usually meant by the

term "language ability," and is the phase which receives special attention in language instruction

The accompanying outline of language outcomes and objectives is a compilation and an adaptation of material from several sources. To develop a complete and perfect outline of language objectives is almost certainly a hopeless task. In spite of certain logical and psychological shortcomings which this outline may possess, it nevertheless gives the teacher helpful suggestions for the identification of useful language skills. The teacher and the student will, of course, wish to revise such an outline from time to time to keep it in line with the best research evidence in the field.

LANGUAGE OUTCOMES AND OBJECTIVES¹

I. Oral Language

A. General Outcomes of Oral Language

1. To form correct habits of articulation, enunciation
2. To assume proper and pleasing body position and mannerisms when speaking
3. To use common courtesy in social groups
4. To speak with feeling, reflecting meaning, thought, and interest
5. To learn how to locate and give information
6. To think while speaking
7. To develop sentence sense
8. To pronounce correctly such words as are used
9. To learn how to acquire new words
10. To speak at a rate suitable to conditions
11. To use voice of suitable clarity and loudness

B. Special Oral Language Outcomes and Situations

1. To relate anecdotes, incidents, etc., interestingly
2. To make necessary simple announcements
3. To participate with ease in conversation
4. To take an active part in arguments, debates, etc.
5. To learn to listen, summarize, and report activities, events, news items, instructions, etc.

¹ Adapted from the following sources (1) Maude McBroom, *The Course of Study in Written Composition for the Elementary Grades* University of Iowa Monographs in Education, First Series, No 10 University of Iowa, Iowa City, 1928 (2) *Iowa State Course of Study for Elementary Schools* State Department of Public Instruction, Des Moines, Iowa (3) H A Greene and H L Ballenger, *Manual of Instructions Iowa Language Abilities Tests* Projected for publication by World Book Co., Yonkers-on-Hudson, N. Y. (4) Paul McKee, *Language in the Elementary School* Houghton Mifflin Co., Boston, 1939.

6. To learn to react properly to social responsibilities, as an introduction, meeting a stranger, etc.
7. To develop ability to assume an active part in school activities, as committee meetings, associations, classroom dramatizations, plays, etc.
8. To learn to take the proper auditor-speaker attitudes
9. To use the telephone properly

II. Written Language

A General Outcomes, Knowledges, and Skills Peculiar to Written Composition

1. To learn to write business and social letters, notes, invitations, etc.
2. To learn to fill in common forms, blanks, etc.
3. To acquire skill in writing notices, announcements, and advertisements, telegrams, etc.
4. To show interest and skill in doing creative writing, such as stories, plays, editorials, diaries, etc.
5. To acquire skill in making outlines from content material
6. To record minutes of meetings, dictations by teacher, etc
7. To acquire skill in evaluating, organizing class and lecture notes
8. To prepare an accurate and comprehensive bibliography

B. Knowledges and Outcomes Peculiar to All Written Work

1. To utilize proper manuscript forms
2. To use proper outline forms
3. To punctuate written work correctly
4. To capitalize correctly

III. Knowledges and Skills Common to Both Oral and Written Language

A Correct Usages

1. To master the most important grammatical usages, as pronouns, verb forms, subject-predicate relationships, redundancy, double negatives, antecedents, etc

B. Rhetorical Skills

1. To develop sentence sense
2. To avoid faults in sentence structure, as useless introductory words, phrases, loose use of connectives, incomplete sentences, etc
3. To organize thoughts and sentences effectively
4. To secure unity and sequence in presentation of thoughts
5. To stimulate interest through use of contrasts, concreteness, variety, simile-metaphors, etc
6. To develop through use a vocabulary rich in color, accuracy, breadth, suitability, variety, etc.

Oral Language Skills. It is apparent from the foregoing catalogue of outcomes that language as a means of

verbal expression appears in two main forms, oral and written. Success in the use of oral language depends in the first place upon the individual's ability to choose, arrange, and enunciate his words in such a manner as to affect his hearers as he intends. In order to guarantee success in the operation of these skills, the pupil must be given training and practice in connected thinking and talking under audience conditions. In this training, emphasis must be placed on the development of a pleasant speaking voice, a gracious attitude, a clear enunciation of words, an avoidance of common language errors, care in the selection of words, a careful selection and organization of ideas, and skill in the clothing of his thoughts in the proper words so that he may affect his hearers as he intends by leading their thinking along prescribed channels.

Written Language Skills. The problem of written language takes a threefold form, although this is not apparent from the outline of outcomes. The first involves the formal or mechanical factors, such as writing, spelling, punctuation, form, and general appearance. The second treats of certain grammatical factors, such as common errors in language form, sentence structure and form, etc. The third is concerned with the more subtle elements of composition, the rhetorical factors involving the questions of choice of words, quality of interest innate in the material, and logical organization of the subject matter both within the sentence and the larger units. In the first two phases of the problem of written language, the factors are more generally uniform in their manner of affecting readers. However, there is greater difficulty in predicting the effect of the third phase on the reader. These mechanical and grammatical elements constitute in a way the raw material of written expression. The rhetorical factors are the results of the manner in which these raw materials are put together. They are the factors which make for appeal, originality, and distinctiveness in written expression. The mechanical and grammatical factors are relatively tangible, objective, and measurable. The rhetorical factors are more elusive, more difficult to identify and to measure, and thus far have eluded the best efforts to measure them objectively.

II. MEASUREMENT AND DIAGNOSIS OF LANGUAGE ABILITIES

Oral Language Scales. An examination of the foregoing outline of language outcomes makes it clear that oral language ability is made up of many related general and specific abilities. It is also equally obvious that from the standpoint of its social utility oral language is extremely important. Yet measurement of oral language abilities is strangely limited. In fact, so far as the writers know, there are no adequate standardized instruments for the measurement of oral composition which will stand inspection in the light of present day criteria. Some progress has been made in the evaluation of techniques for measuring the improvement in oral composition, but thus far no practical way of making the results available to the classroom teacher has been devised. Netzer² made use of electrical recording equipment to secure specimens of oral expression of children in response to various types of stimuli and later prepared three types of oral composition scales from the transcribed material. Experimental use of the scales showed that teachers can be effectively trained to use such scales. However, the practical difficulty of securing recorded specimens of oral expression, and the equally great difficulty of preparing the scales themselves for general use by teachers, have discouraged attempts along these lines.

Diagnosis of Oral Language Disabilities. Considerable progress in the identification of oral language disabilities and in the development of remedial procedures in oral expression has already been made by investigators in the field of speech. It is clear, however, that any classification of speech disorders must necessarily be conditioned by individual points of view. For example, Blanton³ recognizes four fundamental speech disorders: (1) delayed speech, (2) oral inactivities, (3) letter substitutions, and (4) stuttering. On the

² Royal F Netzer, *The Evaluation of a Technique for Measuring Improvement in Oral Composition*. Doctor's Dissertation, University of Iowa, Iowa City, 1937

³ Smiley Blanton, "Problems and Methods in the Correction of Defective Speech" In A M Drummond (Chairman), *Speech Training and Public Speaking for Secondary Schools*. Report of special committee of the National Association of Teachers of Speech. The Century Co., New York, 1925.

other hand, Travis,⁴ in a much more recent and complete discussion of these problems, prefers to group speech disorders under the three following heads: (1) disorders of rhythm in verbal expression, (2) disorders of articulation and vocalization, and (3) disorders of symbolic formulation and expression.

Disorders of Rhythm in Verbal Expression. This group of speech disorders includes stammering and stuttering. Travis, after pointing out their basic similarities, states that "stuttering is characterized by the repetition of sounds, words or phrases, while stammering is characterized by speech blocks." While the number of serious cases of stuttering does not make it a common pedagogical problem, the effect on the individual is so serious that it is important for teachers to have some idea of the nature and extent of this disorder. Careful surveys of school populations indicate that approximately one pupil per hundred will be a stutterer, with the boys far outnumbering the girls in this speech handicap. Apparently there is no very definite or significant relationship between stuttering and the mental ability of the pupil, although Travis points out that other things being equal the mentally defective child, because of immaturity, is somewhat more apt to stutter than is the normal child. In contrast with this, however, he also shows that in at least one large university the stutterers have been found to be superior to the average college student in intelligence. Obviously the selective factors operating to discourage the student who stutters from going to college are significant.

Since the classroom teacher, no matter how great may be his interest in the nature and the causes of stuttering, can do very little about it, the important thing in connection with instruction in oral language is for him to develop the proper understanding of and sympathy for the stutterer's outlook on life.

Disorders of Articulation and Phonation. Normal speech implies the existence of adequate speech equipment in the physical sense capable of responding to the proper stimuli.

⁴ Lee Edward Travis, *Speech Pathology*. D. Appleton-Century Co, Inc., New York, 1931.

Thus it is apparent that the production of speech sounds calls for the most accurate co-ordination of the physical and mental aspects of the speech mechanism. Under this category of disorders of articulation and phonation are classified all of the defects which are found in enunciation and voice production, including delayed development of speech.

Travis points out that in this field there are two types of speech defects, functional defects which are due to bad training, and organic defects which come from injuries or faulty development of the brain or other organs related to speech. Many of this particular class of speech defects arise from such organic difficulties as abnormal development of the tongue, cleft-palate and harelip, abnormal development of the jaws and teeth, adenoids, defective hearing, etc.

The treatment for most of these disorders of articulation and phonation involves medical, mental, hearing, and speech examinations. Since these are generally highly technical in character, they should probably be undertaken only by the trained specialist in each field.

Disorders of Symbolic Formulation and Expression. Travis defines disorders of symbolic formulation and expression as consisting essentially of "a lack of power to execute with ease acts connected with articulated speech and the comprehension of spoken words." The location of these defects is largely a clinical rather than a classroom problem. Accordingly, the teacher, upon the discovery of any cases among his pupils who are unable to articulate or comprehend the spoken word, should immediately refer them to a clinical expert.

Skills Peculiar to Written Language. The catalogue of language outcomes presented on pages 357 and 358 is a reasonably satisfactory classification for the purpose of contrasting the two major types of verbal expression, but it seems inadequate when considered from the point of view of the complete identification and analysis of the specific underlying skills upon which verbal expression depends. For this more exacting purpose a classification based upon such units of language form as the word, the sentence, the paragraph, and the composition unit, and upon certain general mechanical factors, is superior. In order to present a more concrete idea of the types of abilities called into play at each of these

levels of language skill, the accompanying detailed outline is given.

DIAGNOSTIC OUTLINE OF LANGUAGE SKILLS

- A. Words—Skill in the spelling, choice, use, and definition
 1. Spelling—ability to spell certain socially useful words
 - a* Contractions
 - b* Abbreviations
 2. Choice of words
 - a* Same
 - b* Opposite
 - c* Exact word for meaning
 - d* Variety
 - e* Meaningful words
 - f* Minimum number of words
 - g* Semantic variations in meanings
 3. Correct usage
 - a* Verbs
 - b* Pronouns
 - c* Modifiers
 - d* Nouns
 4. Use of dictionary
 - a* Alphabetizing
 - b* Use of guide words
- B. Sentences—Skill in the use, form, structure, and organization
 1. Form
 - a* Complete, coherent, unified
 - b* Variation in beginning
 - c* Variation in length
 2. Kind
 - a* Declarative
 - b* Interrogative
 - c* Exclamatory
 3. Structure
 - a* Simple, compound, complex
 - b* Subject and predicate
 - c* Variety in structure
 - d* Language usage—avoidance of slang and foreign expressions, faulty expressions, double negatives
 4. Organization
 - a* Logical sequence of ideas
 - b* Variety for interest
- C. Paragraphs—Skill in the form, structure, and organization
 1. Form
 - a* Indentation
 - b* Initial and terminal line length
 - c* Length

2. Structure
 - a.* Unity
 - b.* Coherence
3. Organization
 - a.* Outline
 - b.* Logical sequence of ideas
- D. Letter writing—Skill in use of form and mechanics in
 1. Business letters
 2. Social letters
 3. Informal notes
 4. Formal notes
- E. Outline form
 1. Organization
 2. Capitalization
 3. Punctuation
- F. Bibliographical form
 1. Arrangement for unpublished material
 2. Arrangement for published material
 3. Capitalization
 4. Punctuation
- G. General mechanical factors—Skill in control of
 1. Capitalization
 - a.* Initial words in sentences
 - b.* Proper nouns
 - c.* Proper adjectives
 - d.* Titles of honor and respect
 - e.* Important words in titles of stories, articles, etc.
 2. Punctuation
 - a.* End
 - b.* Series
 - c.* Quoted matter
 - d.* Special situations
 3. Margins
 - a.* Top, bottom, sides
 - b.* Indentation
 4. Handwriting
 - a.* Legibility
 - b.* Speed
 5. Abbreviations
 - a.* Titles
 - b.* Other situations
 6. Hyphenations
 - a.* Compound words
 - b.* Ends of lines

In spite of the detail of this outline and the number of specific skills which contribute directly to language ability,

the reader will immediately recognize certain significant weaknesses. Many of the skills are identified only in a very general way. The recognition of choice of words as a significant language skill is approximately equal to stating that addition is an important skill in arithmetic. Much more definite information is necessary before all of the details of a constructive program of language improvement can be developed. Just as it is necessary to identify the socially useful situations and facts, or the most useful words in spelling, it is necessary to identify the skills which have the greatest social usefulness in language situations. Much excellent work has been done on the problem of determining a minimal spelling (writing) vocabulary based on social utility. Similar work must be done from the standpoint of language situations. Until this is accomplished, workers interested in the development of diagnostic exercises in oral and written language must turn to other sources for valid test and drill materials.

Measures of General Merit of Written Composition. The measurement of general merit of written composition, while dating well back into the history of educational measurement, has not responded to efforts to improve it in proportion to the attention it has received. This difficulty comes from the great complexity of the skills involved in producing merit in written language, and from the vagueness with which these skills have been recognized. Historically, the *Hillegas Composition Scale* antedates most other attempts to measure educational products.⁵ Not only has this scale accomplished much good through the stimulation of interest in the more accurate measurement of written composition, but it is still a usable instrument in its present form, the *Thorndike Extension to the Hillegas Scale for the Measurement of Quality in English Composition by Young People*.

Among the more useful of the supplements to the Hillegas scale is the *Nassau County Supplement* developed by Trabue for use in the survey of Nassau County, New York. The first seven of the themes comprising this scale were written

⁵ Milo B. Hillegas, "A Scale for the Measurement of Quality in English Composition by Young People" *Teachers College Record*, 13 331-84, September 1912.

on the topic "What I Should Like to do Next Saturday." The specimens having values of 7.2, 8.0, and 9.0 were selected from compositions published by Thorndike.

The *Willing Scale for Measuring Written Composition* is made up of eight specimens of composition all written on the subject, "An Exciting Experience." Through the definite recognition of the relation of form errors to the general quality of written work this scale increases its usefulness. Its value is also enhanced through the very clear directions for the collection of compositions for survey purposes. An excellent list of interesting topics is also suggested as the basis for the written work. The use of such standardized lists of topics and the control of conditions under which the writing takes place add distinctly to the reliability with which written composition abilities may be measured.

The social importance of letter writing as a phase of written language would seem to justify the more extensive use of letter writing scales such as the *Lewis English Composition Scales*. These scales are standardized for use in grades four and above, as are the majority of such scales. They may be used effectively wherever instructional emphasis is given to writing of any of the following types: order letters, letters of application, simple narrative social letters, simple expository social letters, and simple narration.

Standard Tests in Grammar and Usage. While the measurement of the common grammatical usages is not confined to the field of written language, the very nature of the subject matter itself makes it necessary to measure it in written form. For those who believe that there is a formal as well as a functional aspect of usage, the *Kirby Grammar Test* should have an appeal. This test is intended to measure the pupil's ability to recognize the correct reasons for his choice. The content of the usage exercises is based upon a composite of studies of the typical errors of children.

The *Pressey Diagnostic Tests in English Composition* are useful measures of certain mechanical skills in written language in the junior high school grades. Test A presents twenty-eight exercises in capitalization; Test B calls for responses to thirty punctuation exercises; Test C deals with grammar in thirty exercises of the multiple-choice (one out

of four) type; and Test D presents twenty-four multiple-choice exercises dealing with sentence structure and sentence sense.

Considerable experimentation in attempting to determine the best type of testing technique to use for measuring these mechanical aspects of language indicates that the forms used in the *Pressey Diagnostic Tests in English Composition* are not the most effective and economical. Spaulding has shown that the type of exercise used in Pressey, Test B, Punctuation, is not so reliable per unit of testing time and requires considerably longer to score than does an exercise of the recognition-correction type in which a single punctuation skill is measured.⁶ Stickney showed the outstanding superiority of the recognition-correction form of exercise when used to test usages as compared with the four-choice multiple-choice exercise of the type used in the Pressey tests on grammar and sentence structure. Not only did the recognition-correction type of exercises require much less than half as much time for pupil reaction as the other types, but there was great economy of time in correcting the exercises.⁷

Analytical Measurement of Language Abilities. In addition to the foregoing tests, each of which presents only limited analytical possibilities in the measurement of language, there are three or four others which should be mentioned. In the light of the criteria for diagnostic measurement which have been set up in this volume, most of these tests fall short of being really diagnostic. In fact, it is very doubtful if there are any truly diagnostic tests in the language field. The *Franseen Diagnostic Tests in Language* are diagnostic only to the extent that they identify difficulties dealing with pronouns, verbs, and varied constructions. In spite of this fact they are very useful tests for survey purposes in Grades 3 to 8. The *Language Section of the Stanford Achievement Test* deals with usage only. The *Unit Scales of Attainment in Language* are considerably more comprehensive, dealing with three aspects of language ability: capitalization, punctuation, and usage.

⁶ E. R. Spaulding, *A Critical Study of Two Methods of Testing Punctuation*. Unpublished Master's Thesis, University of Iowa, Iowa City, 1930.

⁷ George E. Stickney, *A Comparison of Two Objective Methods of Testing Language Usage*. Unpublished Master's Thesis, University of Iowa, Iowa City, 1930.

The *Iowa Language Abilities Tests* represent a program of analytical measurement in language which appears to have more than ordinary possibilities. The Primary test is designed and standardized for use in Grades 1, 2, and 3. The following types of abilities are measured filling in forms, conversation, oral composition, telephone conversation, correct usage, recorded composition, miscellaneous social usages, and letter writing. Only parts of filling in forms are attempted in the first grade. Letter writing is tested by recognition in Grades 2 and 3 only. The tests may be administered as group measures in the second semester of the first grade and in the second and third grades. In general, the pupils' reactions are simple and objective. A feature of the test is the use of a single small test booklet for both forms of the test. The column of the directions followed in the Examiner's Manual determines the form of the test which is administered.⁸

The Intermediate test and the Advanced test follow the pattern of the content of the original *Iowa Elementary Language Tests*, which these instruments displace. The Intermediate battery is for use in Grades 4 to 6; the Advanced test is for use in Grades 7 to 10. In appearance and in testing techniques, these tests resemble the *Iowa Silent Reading Tests*.

Illustrations of the techniques of testing word meaning, capitalization, and punctuation are given on page 369.

Test C of the *Iowa Basic-Skills Tests* is a language test of considerable analytical power. The Elementary test is designed for use in Grades 3, 4, and 5 and the Advanced test for Grades 6, 7, and 8. New editions of these tests are issued annually for use in a local state testing program. Following the mid-year program the tests are made available for general use.

III. REMEDIAL INSTRUCTION IN LANGUAGE

Remedial instruction in language will be effective only to the extent that pupils are made aware of the social im-

⁸ H A Greene and Lou A Shepard, *Iowa Language Abilities Test, Primary* Scheduled for immediate publication by World Book Company.

EXCERPTS FROM IOWA LANGUAGE ABILITIES TESTS^a

TEST 2 WORD MEANING

One of the five numbered words in each exercise has almost the same meaning as the first word. One of the words means almost exactly the opposite of the first word. Find the word which means the same as the first word. Note its number. Then fill the answer space under SAME at the right which has the same number as this word. Next find the number of the word which means the opposite of the test word and fill the answer space under OPPOSITE at the right which is numbered the same as this word. Study the samples below.

Summary:

A High 1 slim 2 tall 3 short 4 dark 5 large 6 ^{Opposite} 7 8 9 10 11 12

B Cold 1 sick 2 warm 3 tired 4 chilly 5 lonely 6 7 8 9 10 11 12 ^{Opposite}

TEST 6 CAPITALIZATION

Directions: In some of the following sentences a word is written with a small letter which should begin with a capital letter, or a word is written with a capital which should not be capitalized. Each such word is underlined. Notice the number of this word. Then fill in the answer space at the right which is numbered the same as the word in the sentence which is not written correctly. Some of the sentences are written correctly. If a sentence is correct, fill in the answer space under *N*. The samples are answered correctly. Do the remaining exercises in a similar manner.

SAMPLES: A ¹my ²mother ³came ⁴home ⁵early _____ A ¹ ² ³ ⁴ ⁵

B Did ¹Jim ²take ³the ⁴car? _____ B ¹ ² ³ ⁴ ⁵

C I ¹live ²in ³the ⁴country? _____ C ¹ ² ³ ⁴ ⁵

TEST 7 PUNCTUATION

DIRECTIONS: In each of the following sentences certain words are printed in type like *this*. This means that you are to look for some punctuation mark which may be needed before, within, or after *this* word. In some cases no punctuation is needed. Study each sentence and decide which punctuation mark, if any, is needed. In the answer spaces at the right, fill in the space under the correct punctuation mark to use in the sentence at the place indicated by the word. If you think that no punctuation is needed fill the answer space under *N*. The samples are answered correctly.

SAMPLES **A** We saw an apple on the tree — — — — — **A** 11 11 11 11

B Are you coming with me? — — — — — **B** 11 11 11 11

portance of correct language usage (with the corresponding dissatisfaction with error) and are led to develop a desire to make use of the best forms of expression and to formulate correct habits of usage. Language tests of the analytic types should aid in the developing of a self-critical attitude on the part of the pupil which naturally leads to the desire to acquire correct habits of expression.

Remedial Suggestions on Punctuation and Usage. Specimen types of remedial exercises in language are not presented here for two reasons. In the first place, there are countless excellent practice and drill books in the language field which provide adequate experience in the important

⁹ H A Greene and H L Ballenger, *Iowa Language Abilities Tests* (1) Test 2, Intermediate, (2) Test 6 and Test 7, Advanced Scheduled for immediate publication by World Book Co

TABLE XIV DIAGNOSTIC AND REMEDIAL CHART PUNCTUATION AND USAGE¹⁰

PUNCTUATION

Possible Causes of Low Test Score	Addition of Evidence of Deficiency	Suggested Remedial Treatments
1 Lack of knowledge of the specific punctuation skills	1 Personal check of individual pupils' test paper to determine types of skills missed in the test; observation of daily written work	1 Check the punctuation items missed in the test with textbook and local course of study. Cross check these items with grade placement list of punctuation skills given above. Give individual drill on proof reading tests, emphasizing the specific types of skills missed by the pupil. Drill calling for two definite types of reaction in punctuation is desirable. Certain exercises should emphasize the insertion of the correct punctuation where no punctuation appears. Others should give the pupil experience in critically editing his own or other copy for improper or excessive punctuation.
2 Tendency to over-punctuate	2 High correction for over punctuation in Part III. Analysis of punctuation used in daily written work	2 Use dictation and proof reading drills calling for the elimination of improper or excessive punctuation. Tendencies in this direction will be shown by checking the 15 items listed at the end of the grade list of skills.
3 Lack of a self-critical attitude toward own written work	3 Careless punctuation in daily written preparations in other subjects. Limited ability to note errors in punctuation either in own papers or other written copy	3 Emphasize importance of self criticism of pupil's own daily written work. Drill on proof reading exercises designed to emphasize use of capitals in the types of situations in which the pupil shows weakness.
4 Poor general reading comprehension	4 Low scores on Test A or other reading comprehension tests	4 Drill on sentence comprehension and comprehension of total meaning as suggested in section on Test A.
5 Poorly developed sentence sense	5 Low score on Part II, Test C	5 Carefully explain the various types of sentences and the relation of sentence structure to punctuation. Stress individual practice in writing sentences and punctuating them correctly.
6 Poor vision	6 Observation, nurse's or medical attention	6 Refer to doctor for medical attention. Encourage the pupil to make a special effort to write carefully and to make his punctuation marks distinctly.
7 General carelessness in matters of form in written expression	7 Observation and analysis of characteristics of handwriting and punctuation in daily work	7 Stress continually the essentials of good form in writing. Have pupil do self-critical work on his own papers before submitting them.

¹⁰ Manual of Administration and Interpretation Iowa Every Pupil Basic Skills Tests, pp 74-77 Bureau of Educational Research and Service, University of Iowa, Iowa City, 1939

Usages

1	Poor control over specific usages	1	Observation and checking of daily oral and written expression.	1.	Check pupil's test paper to determine the classes of items missed by the pupil. Check with text and course of study for grade emphasis. Emphasize individual drill on specific points of error. In teaching the correct form, contrast it with the one to be avoided. Sound is important in usage, therefore, supplement written exercises with oral drill. Exercises paralleling sections of diagnostic language tests will be found very helpful.
2	Failure to comprehend the teaching technique	2	Failure to follow directions in recording responses in test, with correspondingly low score on Part V	2	Prepare drill exercises similar to the type of exercise utilized in Part V. Work with individual pupil until he understands the technique.
3	Poor language background	3	Careless and inaccurate usage in all oral and written expression. Poor enunciation and pronunciation.	3	Corrective instruction is the only remedy here. Pick a limited number of important usages and proceed as in No. 1 above.
4	Poor general reading comprehension	4	Erratic responses in Part V. General observation of reading ability in other elementary subjects.	4	Drill on sentence and total meaning comprehension as suggested in section on Test A.
5	Low mental ability	5	Difficulty in following directions, with erratic responses to exercises in Part V, observe difficulty in mastery of common usages. Low mental age and low I Q as revealed by reliable mental examination.	5	Follow general procedure outlined in No. 3 above.
6	Confusion resulting from emphasis on formal rather than functional usages	6	Inaccurate responses on items which may receive instructional emphasis mainly through formal statements of rules.	6	Emphasize individual drills, stressing the establishing of definite habits of correct response to important usages.
7	Careless language habits	7	Inaccurate and erratic responses to items in Part V. Observed carelessness in language habits in informal expression.	7	Emphasize the development of self-critical attitude toward usages. Try to bring to bear social pressure from the group, favoring correct usage. Stimulate the individual pupil to proof-read his own written expression before submitting it.
8	Foreign language influence	8	Use of a foreign language in the home, particularly homes in which two languages are emphasized. Observed foreign accents in pronunciation.	8	See suggestions outlined in No. 1 above.

skill areas. In the second place, the parallel between the desirable types of language drills and the types of exercises used in the tests to reveal the presence or absence of the skills is very close.

One of the most helpful organizations of diagnostic and remedial suggestions was presented in the Manual of Administration and Interpretation of the *1939 Iowa Every-Pupil Basic Skills Tests*. A reproduction of the suggestions for the improvement of punctuation and language usage are presented on pages 370 and 371 as examples of this type of material. Similar suggestions for the improvement of work in spelling, sentence sense, and capitalization are also given in this manual.

Remedial Exercises on Form and Appearance. The social importance of form and appearance in letter writing, one of the most frequent social uses of written composition, places a premium on such skills in this field. Exercises of the following types are suggested as useful drills on letter form.

LETTER FORM EXERCISE

Arrange, capitalize, and punctuate properly the following items which go to make up the heading and salutation of a letter.

gentlemen
236 erie avenue
columbus city mo
july 18 1926
the sanford and morris company

(Use your own home city and address in the heading of the letter)

LETTER FORM EXERCISE

Write as I dictate the following items in the proper form for a business letter

Your home address is 140 Grand Avenue Boulevard.

Your home town is Bluffton, Pennsylvania.

The date is (*give present date*).

You are writing to Williams and Burke, Attorneys-at-law, whose address is Sprague Building, 10th and Ferry Sts, Long Beach, California.

Use the proper salutation (gentlemen) for a letter of this sort.

Remedial Exercises on Sentence Structure. Exercises of the following sort will afford effective remedial drill for children having difficulties in sentence structure.

SENTENCE STRUCTURE EXERCISE ¹¹

In each of the following groups of sentences there are some statements which are not well expressed. Place a cross (X) before each such statement to show that it is not a good sentence.

I asked her the name of the book she was reading

He was glad to leave, for he was tired and sick of the place, for he had made no friends.

She told me the names of her sister and kitty.

Mary had a good position. Which she left.

She is happier than I

Remedial Drill on Choice of Words. A great deal of difficulty is encountered by youthful (and adult) writers in the choice of words. Such errors as "He does his work good" are not uncommon. In many cases it is a matter of choosing a word with a more exact shade of meaning, and in many cases it is a matter of knowing the form of a given word to use. This is particularly true in the field of adjectives and attributive nouns. Exercises such as the following, giving drill in choosing the correct compared form of certain adjectives, will be found helpful in cases in which specific diagnosis reveals this type of weakness.

EXERCISE IN CHOICE OF WORDS ¹²

Read the following exercises and fill in the missing word needed to complete the meaning of the sentence. In each the missing word in an exercise is a form of the underlined word in the sentence.

1. New York is a large city. It is _____ than Chicago, in fact it is the _____ city in the United States.
2. I want to buy a good fountain pen. I want a _____ one than that. Show me the _____ one you have.
3. John is a mischievous boy. He is _____ than Bob. He is the _____ boy in school.
4. This is a bad storm, much _____ than the one we had last week, but not the _____ one this year.

IV. IMPORTANCE OF MEASUREMENT IN SPELLING

Social and Educational Significance of Spelling. The importance of correct spelling in the written communica-

¹¹ Adapted from Course of Study in Language, University Elementary School, University of Iowa, Iowa City

¹² Ibid.

tion of ideas is quite generally recognized. Applicants for positions have often failed to receive employment because of incorrect spelling of words in their letters of application. Business and social status is frequently determined to a large measure by a person's mastery or lack of mastery in this specific skill. Spelling, because of its social significance and its tool value in connection with later school progress, is so important that educators in general are unwilling to depend upon the incidental teaching of it for the development of the required skill. Spelling is recognized as one of the subject fields in which the learning is specific. The child does not just learn spelling, but he learns to spell specific words. He may master a definite method of learning to spell, but the words he learns to spell are mastered as a result of a definite application of effort and attention.

Purposes of Teaching Spelling. The objectives of the teaching of spelling as stated below¹⁸ appear to contain most of the essentials.

1. To make automatic the accepted sequence of letters most commonly needed for expression of thought in writing.
2. To develop the meaning and use of words to be spelled. The development of the meaning and use of words may involve the meaning and uses given in the dictionary, but it is preferable to clarify and to build up the meaning and uses upon the basis of the child's own experience.
3. To develop what is termed a 'spelling consciousness,' i.e., the ability to recognize almost instantly the correct and incorrect spelling of words.
4. To develop a 'spelling conscience.' This 'spelling conscience' may be referred to as an ardent purpose or desire to spell correctly, or as an ideal of correct spelling. This conscience is annoyed by incorrect spelling and is satisfied only with correct spelling.
5. To develop a technique for the study of spelling. This technique involves the application of an effective method of learning how to attack and master the sequence of letters in the given word, the method of diagnosing sources of error in the spelling of specific words, the knowledge of how to use the dictionary in finding the pronunciation, meaning, and correct spelling of unfamiliar words, the knowledge of what to do when in doubt concerning the spelling of a word, and the application of a few inductive rules governing the correct spelling of words.

¹⁸ "Spelling." *The Nation at Work on the Public School Curriculum*. Fourth Yearbook of the Department of Superintendence, Chapter VI, pp. 126-27. National Education Association, Washington, D. C., 1926.

Measurable Qualities in Spelling. Although it is obviously impossible to provide objective measures for all of the specified outcomes of instruction in spelling, such a list of outcomes furnishes the best basis for the identification of the measurable qualities in the field. The ability to spell in list or in context those words which are most commonly needed in written expression may be measured quite satisfactorily by means of samplings of words taken from vocabulary lists of known social importance. The ability to recognize the correct or incorrect spellings of socially useful words is quite readily measured by means of proof-reading tests. This same type of measure may be used within certain limits in the determination of the development of a "spelling conscience." It is doubtful whether there are any existing tests suitable for the measurement of the acquisition of new word meanings. This is also true of the development of good study habits in spelling. A close observation of the pupil's daily work in spelling probably affords the best check on his use of proper study habits. Good habits of work in spelling, as in other subjects, are usually revealed indirectly in the results.

Early in the consideration of the problems of measurement in spelling, two aspects of the pupil's accomplishment in this subject should be pointed out. In the first place, there is the problem of determining the pupil's present spelling ability. The second aspect of the problem concerns the measurement of progress. This is expressed in terms of the improvement which the pupil makes, under instruction, in the mastery of the specific spelling vocabulary on which he is at work. Thus the determination of the child's spelling ability should be undertaken prior to study on the specific list of words. At the end of the teaching process a second measure is taken. This affords an indication of how much the child has progressed in his mastery of the selected spelling vocabulary.

Systematic Sampling of Words. The introduction of scientific methods in education in recent years has resulted in many investigations into the scope and character of spelling lists. Studies such as those by Anderson, Ayres, Fitzgerald, O'Shea, Thorndike, and Horn seem to warrant the conclusion that approximately 4000 carefully selected words

would be an appropriate number for the basic spelling list for elementary schools. Furthermore, these studies have proved of great value in selecting the word-lists to be included in spelling texts, tests, and scales. It is quite obvious that the words most commonly used in the written language activities of adults and children should receive the major emphasis in a spelling course of study. To teach pupils to spell words that they will very rarely be called upon to spell either in or outside of the school is clearly a waste of time. Such words are best left to incidental learning or to the responsibility of each person as the need for their use arises.

V. CONSTRUCTION OF SPELLING TESTS

In the construction of spelling tests there are at least four problems which require careful consideration. These are discussed briefly in turn.

What Words. One of the first problems in the construction of a spelling test is that of selecting the words to be included. The values of spelling are almost entirely specific, and lie in the ability of the pupil to spell words which are actually used and are most certain to be used. It is important, therefore, that the content for a test should be sampled from those words that are and will be ultimately of maximal usefulness to the pupil.

Theoretically, spelling lists may be taken from the writing vocabularies of children, the writing vocabularies of adults, or from words common to both. Breed suggests the following ¹⁴

1. That the most important constituent of the minimal spelling vocabulary is a list of words with relatively high frequency in the written discourse of both children and adults
2. That words of especially high frequency in the usage of children should be included in the minimal list, regardless of adult usage.
3. That words of especially high frequency in the usage of adults should be included in the minimal list, regardless of the usage of children, but care should be taken to allocate these words to the upper elementary grades

¹⁴ F. S. Breed, "What Words Should Children Be Taught to Spell?" *Elementary School Journal*, 26 292-306, December 1925

Among the word lists which have been widely used in the construction of spelling tests is one by Anderson,¹⁵ the Thorndike *Teacher's Word Book*,¹⁶ and the Horn *Basic Writing Vocabulary*.¹⁷ Anderson's list was one of the first to be based on an extensive word count. Thorndike's list contains ten thousand words which were found to occur most frequently in a count of several million words taken from many sources. Horn's list includes ten thousand words chosen from varied types of adult writing. The words are classified on the basis of frequency, and each word frequency is compared with that given in other vocabulary studies. This study took into account all previous spelling vocabularies, and as a result has greatly influenced the content of recent spelling texts. Only 3009 of the ten thousand words in this writing vocabulary were designated by the author as basic for elementary school spelling lists.

The *Iowa Spelling Scales*, representative of another source of information on what words to include in a spelling test, were based on the 2977 words found by Anderson to be most frequently used in written correspondence.

Teachers who are using spelling texts made up of word lists of unknown social importance will find such sources of great value in selecting valid content for their own tests. Words to comprise a spelling test should, of course, be among those comprising the list studied by the pupils. The most valid types of spelling words on which to test a pupil are also those words which have relatively high social usage. Thus a cross-check of the words common to the local spelling text and to the *Iowa Spelling Scale* (or other similarly developed scales) will reveal the high social-frequency words which the pupils have studied and will at the same time give the teacher a measure of the relative difficulty of the words from their values in the scale itself. Thus the teacher may construct his own valid test on words of known difficulty.

¹⁵ W. N. Anderson, *Determination of a Spelling Vocabulary Based upon Written Correspondence* University of Iowa Studies in Education, Vol II, No. 1 University of Iowa, Iowa City, 1917

¹⁶ E. L. Thorndike, *The Teacher's Word Book* Teachers College, Columbia University, New York, 1921

¹⁷ Ernest Horn, *A Basic Writing Vocabulary* University of Iowa Monographs in Education First Series, No. 4 University of Iowa, Iowa City, 1926.

How Difficult Words. It is well known that some words are more difficult than others, i.e., some words are more frequently misspelled than others. If words are selected at random from any of the lists indicated above, some of them will be easy and some relatively difficult. If equal credit were given for each word thus selected, an error would be introduced into the estimate of the pupil's spelling efficiency. Words for a test should be selected in terms of their known difficulty. The words of the *Ayres* and the *Iowa Spelling Scales* have been so classified by having the words spelled by large numbers of children, and the relative difficulty of each word determined by the percentage of correct spellings of each word. The words to be included in the test for any grade should be adapted if possible to the ability of the group to be tested. Classes of average ability appear to respond best to words of approximately 50 percent difficulty.¹⁸ On the other hand, if the test is to be given over a wide spread of ability, words ranging from 14 to 86 percent standard accuracy with a mean of fifty percent tend to give a distribution more closely approximating the normal frequency curve, with the pupils grouped more closely around the mean. In general, it is probably safe to say that the words to be included in a test for any grade should be those on which there are from 40 to 70 percent misspellings. Tests made up of such words will give a reliable measure of spelling ability, since the words will not be so easy that there will be many perfect scores or so difficult that there will be many low scores.

How Many Words. The purpose the test is to serve will determine the number of words to use. For survey purposes a list of 25 words will probably be sufficient to determine the status of spelling efficiency for a school system. To be sure, the ability to spell one word is separate and distinct from the ability to spell other words. It would seem necessary, therefore, to subject a pupil to the 1000 words of the *Ayres* list or to several hundred of the *Iowa* lists in order to secure a reliable measure of his ability to spell the most

¹⁸ Walter W. Cook, *The Measurement of General Spelling Ability Involving Controlled Comparisons between Techniques*. University of Iowa Studies in Education, Vol. VI, No. 6. University of Iowa, Iowa City, 1932.

commonly used words. However, the procedure of sampling applies to the testing of spelling as in all other testing. While 25 words is possibly a sufficient number for survey purposes, a larger number is needed to reveal the spelling ability of individual pupils. On the whole it appears that a minimum of 100 words should probably be used for individual testing purposes in spelling. Possibly 50 words are not too many to use for the measurement of general class accomplishment.

How Given. The question of the form in which spelling words should be presented for testing purposes has called forth much debate in the past. It has also been subjected to experimental study with results which are not too conclusive when considered in the light of practical classroom procedures. Horn summarizes the evidence on this question as follows:¹⁹

Written tests are to be preferred to oral tests . . . Recall tests are superior to and more difficult than recognition tests. The evidence indicates that the most valid and economical test is the modified sentence recall form, in which the person giving the test pronounces each word, uses it in an oral sentence, and pronounces it again. The word is then written by the students.

VI. DIAGNOSIS AND REMEDIATION OF SPELLING DISABILITIES

Spelling tests and scales afford useful sources of material which may be used to determine both the pupil's present status in spelling and his growth in accomplishment as a result of a period of instruction. If scales based on a sound philosophy of subject-matter content are used, they provide the most effective materials for the identification of the spelling difficulties of individual pupils. Samplings from scales used as tests give the teacher an objective basis for the study of these personal difficulties through the accumulation of individual lists of words which are sources of trouble.

Cumulative Personal Lists. To a large extent remedial procedures in spelling may be undertaken directly in connection with teaching. The words misspelled by pupils in

¹⁹ Ernest Horn, "Spelling" *Encyclopedia of Educational Research*, p. 1179. The Macmillan Co., New York, 1941.

their spelling lessons and tests are obviously the words to which they should give special attention. Each pupil should be encouraged to keep an individual list of such words and should be stimulated to master them. Occasional spelling periods should be put aside for studying and testing these individual lists. If such lists are properly utilized, each pupil will come to regard his individual "demon" list as a means for eliminating spelling deficiencies rather than as a form of punishment for misspellings.

Analysis of Written Work. Every pupil's written work in all subjects should be carefully checked for spelling errors. Such a list of misspellings in written work should be kept by every pupil, and he should realize that he is to be held responsible for the mastery of these troublesome words. The important thing is that the learning situation be so manipulated that the pupil will want to learn to spell and to feel the need for learning the meaning and spelling of words that are pertinent to his written work. The teacher must see to it that pupils develop a sense of pride in correct spelling and dissatisfaction over incorrect spelling. If this has been accomplished, the majority of pupils will soon assume a new responsibility for spelling correctly the words used in their written work. They will check their spelling before turning in their papers. They will habitually appeal to authority when in doubt, which implies the development of desirable habits of using the dictionary.

Probably nothing short of the consistent and united effort of all teachers will make possible the realization of a high degree of vocabulary skill by all pupils. The tendency to integrate the language arts and the other subjects of the elementary school is a very promising lead in connection with the problem of vocabulary enrichment.

Individual Pupil Diagnosis. The discovery from the results of a spelling test that a pupil is below the norm in spelling ability may be of considerable value, but it falls far short of its real function unless it reveals to the pupil the particular weaknesses which resulted in his low score. The following items of information procurable through observation and measurement are invaluable in diagnosing individual pupil disabilities and should be used as much as

possible in connection with the analysis of the pupils' spelling habits: (1) intelligence quotient, (2) spelling marks, (3) reading marks, (4) writing marks, (5) attendance data, (6) visual-defects data, (7) auditory-defects data, (8) speech data, (9) general health data, (10) personality characteristics—industry, aggressiveness, independence, attentiveness, exactness.

Many investigators of spelling disabilities have abandoned the procedure of deducing the causes of spelling difficulties from an analysis of errors and are now devoting their time and energies to studying the work habits of pupils by means of careful observation and tests.

Study Techniques. One of the chief causes of poor spelling achievements lies in the failure of the pupil to utilize an adequate method of study. Frequently this is not the fault of the pupil. Poor study technique upon the part of the pupil may be detected through careful observation and testing. Spelling consists in forming correct and fixed associations "between the successive letters of a word and between the word and its meaning"²⁰ The fixing of these associations depends upon the use of effective techniques of habit formation. Observation should reveal whether or not the pupil is systematic in his attack upon new words. Does he center his attention upon it, does he practice visualizing it, i.e., closing his eyes and attempting to see the word, does he spell the word softly to himself, does he pronounce it syllable by syllable and try to think how each syllable looks; does he compare the word with the correct written copy; does he watch for silent letters, double letters, different vowels having the same sound, and for hard groups of letters; does he rewrite the word repeatedly until he has it completely mastered, does he develop the meaning for the word by using the dictionary and by using it in different sentences?

Pronunciation and Articulation. Another significant cause of spelling disability is imperfect pronunciation and articulation of words. The contributing factors may be imperfect hearing or improper pronunciation and enunciation of the

²⁰ Frank N. Freeman, *Psychology of the Common Branches*. Houghton Mifflin Co., Boston, 1916

words by parents, teachers, and pupils. Frequently words are misspelled because they are spelled as they are heard or pronounced. It is interesting in this connection to note that improper pronunciation by the pupil himself is even more likely to cause errors in spelling than such mispronunciation by the teacher in presenting the word either in teaching or testing. Children should pronounce the words to themselves in learning to spell them. It is therefore very important that they form proper habits of pronunciation and enunciation.

Association of Sounds and Syllables. Syllabication is undoubtedly a factor influencing proper enunciation and pronunciation of words and the ease with which they are learned. The syllable is a natural unit in vocalization and visualization, and dividing a word into syllables does not in any way interfere with the mental image to be secured of the whole word. Pupils are aided by securing a clear visual image of the word by syllables, and by sounding the word by syllables.

Grouping words which have similar phonetic elements is also of value in learning to spell them, particularly for pupils beginning their work in spelling and learning the sounds of letters and common phonograms.

Spelling Consciousness. The lack of spelling consciousness constitutes one of the serious causes for poor spelling ability, particularly in written composition. Awareness of correct spelling should carry over into all normal writing activities of school and life. The pupil should learn to recognize promptly correct and incorrect spellings of words. Clear understanding of word meanings and of correct pronunciation is valuable in the development of spelling consciousness.

Spelling Conscience. Closely related to spelling consciousness is the pride in spelling ability, or spelling conscience, which results in deep dissatisfaction in a pupil every time he misspells a word. A realization of the import of incorrect spelling will make the pupil purposeful in his spelling.

Remedial Work in Spelling. Poor spelling is due to faulty or inadequately formed associations. Basically all

spellers, good or bad, learn in the same way — through association. The main difference between the able and the poor speller lies in the study-technique, personality characteristics, and emphasis he gives to the subject.

Tidyman suggests the following procedure in diagnosing and treating problem cases in spelling: ²¹

1. Give a standard spelling test to discover the amount of deficiency.
Compare with achievement in other subjects
2. Give an intelligence test to discover general mental capacity.
3. Test for defects of hearing and vision.
4. Give reading test.
5. Give test of spelling consciousness to show whether mistakes are due to carelessness or ignorance of the word.
6. Collect misspellings from spelling tests and written work, and classify them according to types of errors.
7. Get as much information as possible about the pupil's pedagogical history, especially methods of beginning reading, knowledge of meanings of words, knowledge of phonics, pronunciation and articulation, motor coordination in writing, and emotional attitude toward spelling.
8. From above, assemble probable causes of difficulty in spelling, and adopt appropriate remedial measures, such as the following
 - (a) Systematic word study. Early training may have been inadequate.
 - (b) Exercises in visualization
 - (c) Drill upon particular types of spelling errors.
 - (d) Phonics drills
 - (e) Removal of physical defects
 - (f) Develop confidence through successful effort.

Horn presents 41 principles of teaching spelling, as derived from scientific investigation.²² His principle No. 40 is stated as follows: "It is important that each pupil be taught *how to learn* to spell." The following rules are suggested by him as designed to embody the conclusions of various experiments in economy of learning. They are in such form

²¹ Willard F Tidyman, In William H Burton (Editor), *The Supervision of Elementary Subjects* D Appleton and Co, New York, 1929

²² Ernest Horn, "Principles of Method in Teaching Spelling as Derived from Scientific Investigation" *Fourth Report of Committee on Economy of Time in Education*. Eighteenth Yearbook of the National Society for the Study of Education, Part II, Chapter IV, pp 52-77 Public School Publishing Co., Bloomington, Ill., 1919.

that they may be used to great advantage by both the teacher and the pupil.

HOW TO LEARN TO SPELL A WORD

1. The first step in learning to spell a word is to pronounce it correctly
If you do not know how to pronounce a word, look up the pronunciation in the dictionary. When you are certain that you know how a word is pronounced, pronounce it, enunciating each syllable distinctly and looking closely at each syllable as you say it
2. Close your eyes and try to recall how the word looks, syllable by syllable, as you pronounce it in a whisper. In pronouncing the word be sure to enunciate the syllables carefully.
3. Open your eyes to make sure that you were able to recall the correct spelling.
4. Look at the word again, enunciating the syllables distinctly.
5. Recall again, with closed eyes, how the word looked.
6. Check again with the correct form. This recall (as in 2 and 5) should be repeated at least three times, and oftener if you have difficulty in recalling the correct form of the word
7. When you feel sure that you have learned the word, write it without looking at the book, and then check with the correct form
8. Repeat this two or more times without looking either at the book or at your previous attempts
9. If you miss the word on either of these trials, you should copy it in your spelling notebook, since it probably is especially difficult for you.

VII. IMPORTANCE OF MEASUREMENT IN HANDWRITING

In spite of the growing popularity and use of mechanical means for writing both in school and out, it is probably true that handwriting will still continue to be the major means of written communication. If it is to serve as an adequate aid in social and business communication, handwriting must be easily read, neat and pleasing in appearance, and of such form that it can be produced under normal conditions with a fair degree of speed.

Objectives of Handwriting Instruction. A practical and concise summary of the objectives of instruction in handwriting is given below:²³

²³ "Handwriting" *The Nation at Work on the Public School Curriculum* Fourth Yearbook of the Department of Superintendence, Chapter V, pp. 113-14. National Education Association, Washington, D C, 1926

OUTCOMES OF HANDWRITING INSTRUCTION

1. To develop sufficient skill to enable pupils to write easily, legibly, and rapidly enough to meet present needs and social requirements.
2. To equip the child with methods of work so that he will attack his writing problems intelligently
3. To diagnose individual writing difficulties.
4. To aid the child to recognize and make use of his peculiar learning capacities.
5. To provide experiences which will tend to develop in the child more power to direct his own practice, and more ability to judge whether or not he is succeeding in that practice
6. To provide the means for each individual to progress at his best rate.
7. To develop an appreciation of the relationship between correct body adjustment and an efficient writing production
8. To secure acceptable and customary arrangement and form for written work (margins, spacing, etc)
9. To develop a social urge to use the skill attained in all writing situations
10. To train pupils to be able, at the end of the sixth grade, to write quality 60 (Ayres Scale) or better, and at the rate of 70 letters per minute or better

Measurable Qualities in Handwriting. Writing involves a very exact type of visual-muscular coordination which must be developed to a high degree if the product is to possess legibility, speed of production, and æsthetic quality. Some difficulty has been encountered in the measurement of certain of the elements of good writing, particularly from the point of view of analysis and diagnosis. The available writing scales, however, have done much to establish for the pupil and teacher rather definite standards or ideas of what constitutes an acceptable product as well as to make both more and more sensitive to handwriting faults and needs.

The measurement of handwriting quality in its refined form is concerned with two factors: (1) quality, or degree of legibility, and (2) speed, or the quantity of writing produced in a given unit of time.

Quality. In modern practice the quality of handwriting is usually determined by comparing a sample of the pupil's handwriting with specimens in a standard scale. While this method of evaluating handwriting specimens is somewhat subjective, experience shows that considerable skill and objectivity can be developed through training in the use of

such scales. At one time the measurement of handwriting simply involved the comparison of the script produced with the copybook sample. This resulted in over-emphasis on the shape and shading of letters and in the formation of beautifully engraved lines. Rate and quality were not the objectives of writing instruction or of measurement under those conditions.

The essentials of quality in writing are measurable within reasonable limits. A number of scales have been developed for use in measuring quality, but they differ greatly in the number of elements of quality measured and in the numerical designation of each quality difference. Therefore, it is difficult to compare the results secured from the use of one scale with those secured from another. The accompanying table summarizes norms of handwriting quality and rate derived from a number of different scales.

TABLE XV
GRADE STANDARDS IN HANDWRITING ²⁴

Grade	Form or Quality			Speed
	Ayres Scale	Thorndike Scale	Freeman Scale	Letters per Minute
II	35	8 5	11	30
III	39	9 3	12 5	44
IV	46	10 2	14 5	51
V	50	11 0	16	60
VI	57	11 9	18	63
VII	62	12 7	20	68
VIII	66	13 5	21	73

Freeman later proposed the quality and rate standards for the *Ayres Scale* as shown in Table XVI. It will be noted that standards are not given for Grades 7 and 8. Freeman assumes that formal penmanship will be discontinued after the sixth grade.

²⁴ Frank N Freeman, "Principles of Method in Teaching Writing as Derived from Scientific Investigation" *Fourth Report of Committee on Economy of Time in Education* Eighteenth Yearbook of the National Society for the Study of Education, Part II, Chapter II, pp 11-25 Public School Publishing Co., Bloomington, Ill., 1919

TABLE XVI
HANDWRITING QUALITY AND SPEED STANDARDS²⁵

Grades	II	III	IV	V	VI
Quality on Ayres Scale	35	45	50	55	60
Rate in Letters per Minute	30	40	50	60	70

The standards for quality as shown in Tables XV and XVI are considered by many to be sufficiently high. The quality values assigned for the various grades are based upon the median performance of many school children. Therefore, these values indicate the quality of handwriting which exists; they do not necessarily indicate the quality which should obtain.

Koos investigated the quality of adult handwriting and also the opinions of adults concerning a satisfactory quality of handwriting. He reached the following conclusions on the basis of his findings²⁶

The fact that some (pupils) will go into pursuits demanding a quality better than 60 should not be offered as a justification for requiring all pupils to attain that better quality. Since all should be required to learn to write as well as 60 for purely social use, to train pupils to write this quality is the task of general education, to teach some who are going into commercial or other vocations requiring a higher quality to write this better quality is the task not of general but of vocational education. . . In the light of these facts it is difficult to see why . . . a pupil should be required to spend the time necessary to learn to write better than the quality of 60. There is even considerable justification for setting the ultimate standard at 50.

Rate. The rate at which pupils write is of considerable importance. The person who is able to write more rapidly than others and with approximately the same quality has an obvious advantage in the field of written expression, provided, of course, that ideas come to him as rapidly as he is

²⁵ Frank N. Freeman and M. L. Dougherty, *How to Teach Handwriting* Houghton Mifflin Co., Boston, 1923.

²⁶ L. V. Koos, "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools" *Elementary School Journal*, 18 423-46, February 1918.

able to transcribe them. The measurement of rate in writing is much less difficult than the measurement of quality. Rate of writing can be measured most conveniently by asking pupils to write, within carefully controlled time limits, selections from standardized copy. If the pupils all write from the same selection and if they have all thoroughly memorized it, the number of letters each pupil writes in the time allotted can easily be computed as the pupil's rate score. Table XVI indicates that pupils at the end of the second grade should write at the rate of 30 letters per minute and increase their speed to 40 letters per minute at the end of the third grade and to a rate of 70 letters per minute at the end of the sixth grade.

VIII. MEASUREMENT OF HANDWRITING ABILITY

Methods of Measurement. Measurement of the quality of handwriting and the rate at which it is produced is accomplished by the evaluation of handwriting specimens secured under standard conditions. Accordingly the first step in the process of measuring handwriting is that of securing these specimens under controlled conditions.

Securing Handwriting Specimens. Three factors appear to affect the conditions under which handwriting specimens for scaling are secured. The character of the copy which the pupils are called upon to write may significantly influence their reactions. It is usually better, at least in the lower grades, to make use of some simple sentence or paragraph which the children have memorized in previous connections, such as "Mary had a little lamb" or some other equally familiar nursery rhyme. The sentence, "A quick brown fox jumps over the lazy dog" has been used on numerous occasions. The chief merit of this sentence lies in the fact that it contains all of the letters of the alphabet. Whatever sample is used should be simple and easily understood, so that the children will not be unduly affected by spelling and vocabulary difficulties. To guard against lapses in memory, it is a good practice to write the copy on the blackboard two or three days in advance of the tests where it can be easily

seen and studied both before and during the collection of the writing specimens.

The instructions which are given to the pupils may also influence the quality and rate of their writing. Therefore, care should be exercised to use very precise directions. The use of this statement in the instructions to the children is recommended: "Write as well as you can and as rapidly as you can."

The time allowance for the writing of the specimens is a third factor which must be considered in the collection of writing specimens. In the standardization of his scale Ayres used the first four sentences of Lincoln's Gettysburg Address and allowed each child two minutes in which to copy as much of this material as possible. Since that time it has become a rather typical practice to allow a two-minute period for the writing of such samples.

The teacher who is inexperienced in the administration of such a test may find the following directions helpful. When the children are all ready, having been provided with paper, and pen and ink or pencil, depending on the grade and the course of study, they should be given a few simple directions. The following are suggested "Write as well as you can at your usual speed, using the following copy: 'Mary had a little lamb' (or some other selected copy). Write the copy over and over until I give the command 'Stop.' When I say 'Stop' you should stop even though you are in the middle of a letter." After these directions have been given the teacher may say, "All in position. Dip your pens. Pens up. Begin." At the expiration of two minutes the command "Stop" should be given and the pupils asked to place their pens on their desks.

Securing Rate Scores. Rate of handwriting is expressed in terms of the number of letters written per minute. This is determined by counting the total number of letters written by each pupil and dividing this by the number of minutes the pupils were allowed to write.

Securing Quality Scores. The quality of the handwriting specimen which is being measured is determined by moving it along the scale until a specimen is found which closely

matches it in quality. The quality value of the scale sample is then assigned as the quality of the sample of the pupil's handwriting.

Accuracy in Measurement of Handwriting. Skill in the evaluation of handwriting specimens requires a thorough understanding of the scale to be used. It is desirable therefore for the teacher, prior to any attempt to use the scale for the measurement of handwriting quality, to study carefully the scale itself, the directions for its use, the norms, and the specific functions which the particular scale is expected to perform. The accurate and reasonably objective rating of handwriting samples on a scale requires considerable skill which experience shows can be developed through practice. For this purpose standard sets of writing samples of known quality are very useful.

Merit Scales. Handwriting scales may be divided into two groups: (1) general merit scales and (2) analytical and diagnostic charts and scales. The choice of a scale depends upon the purpose it is to serve.

The *Thorndike Scale* was the first writing scale to be devised. This scale is designed for Grades 5 to 8 inclusive and consists of a series of specimens of handwriting so arranged that they increase in order of merit from a quality of 4 units above zero to one of 18. Its purpose is to aid teachers in grading handwriting for "general merit" on the basis of three characteristics: beauty, legibility, and character.

*The Ayres Handwriting Scale*²⁷ was the next scale to be devised and it differed from the *Thorndike Scale* in that it was standardized on the basis of legibility. Legibility, the unit of measurement used, was determined by the speed and ease with which samples of handwriting were read by a number of trained and competent judges. The first edition (*Three-Slant Edition*) contained three styles of handwriting: slant, semi-slant, and vertical. This edition contained eight samples of handwriting for each style. The *Gettysburg Edition*, which appeared later, contained only one style of handwriting—the accepted moderate-slant style. Because

²⁷ Leonard P. Ayres, *Scale for Measuring the Quality of Handwriting of School Children*. Bulletin No. 113. Division of Education, Russell Sage Foundation, New York, 1912.

of its convenient form this scale has been one of the most widely used handwriting scales.

The specimens of handwriting for both the Thorndike and Ayres Scales, while presumably sampled from all levels of the school, were taken largely from the upper grade levels. This has made it difficult to measure reliably the writing of the lower grades.

The American Handwriting Scale developed by Paul V. West is one of the most recent and comprehensive of the general merit scales. Among a number of distinctive features are at least two which deserve special mention: (1) A separate scale is provided for each grade from two to eight; (2) The samples have been scaled for both quality and rate, the poorer samples being written at a slower rate and the better samples being the ones written at a more rapid rate. The existence of the separate scales for Grades 2 to 8, inclusive, permits a somewhat more exact evaluation of quality of writing in its relation to grade location.

*The Conard Manuscript Writing Standards*²⁸ are composed of two scales for the rating of manuscript writing. Two separate scales, one for the rating of pencil forms and the other for the rating of pen work, are available.

Legibility of numbers in arithmetic work may be measured by means of the *Knight-McClure Arithmetic Neatness Scale*. Carelessness in the writing of numbers both in and out of school is probably of more importance than would appear at first thought, since in this case contextual relationships do not particularly help in deciphering the meaning of the numerals if they are not clear. The scale itself is made of ten specimens of arithmetic work, stressing numerals ranging in quality from zero to nine. Teachers of arithmetic should find this scale useful.

IX. DIAGNOSIS AND REMEDIATION OF HANDWRITING

Diagnostic Measures. Instruments for the identification of specific faults in handwriting are of two general types: (1) analytical scales and (2) score cards.

²⁸ Edith U. Conard, "Manuscript Writing Standards." *Teachers College Record*, 30 669-80, April 1929

STANDARD SCORE CARD FOR MEASURING HANDWRITING ²⁰

Pupil Age Date
 Grade School
 Sample Number Teacher

Sample	Per- fect Score	Month									
		1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
1 Heaviness . .	3										
2 Slant	5										
Uniformity											
Mixed											
3 Size	7										
Uniformity											
Too large											
Too small											
4 Alignment	8										
5 Spacing of lines	9										
Uniformity											
Too close											
Too far apart											
6 Spacing of words	11							..			
Uniformity											
Too close											
Too far apart											
7 Spacing of letters	18										
Uniformity											
Too close											
Too far apart											
8 Neatness	13	.	.								
Blotches											
Carelessness											
9 Formation of letters	(26)										
General form	8										
Smoothness	6										
Letters not closed	5										
Parts omitted	5		.	.							
Parts added	2										
TOTAL SCORE	100										

²⁰ C T Gray, *A Score Card for Measuring Handwriting*. University of Texas, Bulletin No 37. University of Texas, Austin, July 1915.

*The Freeman Chart for Diagnosing Faults in Handwriting*³⁰ is virtually five scales in one. Each scale is designed to reveal whether the pupil's writing specimen violates one or more of the five essential characteristics of good handwriting. These traits are: (1) uniformity of slant, (2) uniformity of alignment, (3) quality of line, (4) letter formation, and (5) spacing. Each scale shows three levels of quality for the trait with which it deals—excellent, mediocre, and poor. This scale is valuable because it enables both teacher and pupil to discover the weaknesses in handwriting which are in need of special treatment.

Gray's Score Card represents a very important device for analyzing deficiencies in handwriting. This card is reproduced in an accompanying illustration.

Diagnosis by Analysis. Improvement in handwriting instruction depends to a large degree upon the teacher's knowledge of the elements which make for quality in the product, and the use of instruments which are adequate to reveal significant differences in quality. Inferior products of handwriting instruction may be due to lack of skill or mastery in many different phases of the writing act. *Freeman's Chart for Diagnosing Faults in Handwriting* meets this need for securing separate measures of the several aspects of handwriting performance. This scale may be used to measure the whole class, but it is most effective when used to diagnose the writing of pupils who rank conspicuously below the grade norm as revealed through use of some general merit scale.

The following list of handwriting defects and their causes should be useful to the classroom teacher.

ANALYSIS OF DEFECTS IN HANDWRITING AND THEIR CAUSES³¹

Defect	Causes
1. Too much slant	(1) Writing arm too near body (2) Thumb too stiff (3) Point of nib too far from fingers (4) Paper in wrong position (5) Stroke in wrong direction

³⁰ Frank N. Freeman, *The Teaching of Handwriting* Houghton Mifflin Co., Boston, 1915.

³¹ Freeman, *op cit.*

- | | | |
|---|-----------------------|---------------------------------------------------------------------------------------------------------------------------------|
| 2 | Writing too straight | (1) Arm too far from body
(2) Fingers too near nib
(3) Index finger alone guiding pen
(4) Incorrect position of paper |
| 3 | Writing too heavy | (1) Index finger pressing too heavily
(2) Using wrong pen
(3) Penholder too small diameter |
| 4 | Writing too light | (1) Pen held too obliquely or too straight
(2) Eyelet of pen turned side
(3) Penholder too large diameter |
| 5 | Writing too angular | (1) Thumb too stiff
(2) Penholder too lightly held
(3) Movement too slow |
| 6 | Writing too irregular | (1) Lack of freedom of movement
(2) Movement of hand too slow
(3) Pen gripping
(4) Incorrect or uncomfortable position |
| 7 | Spacing too wide | (1) Pen progresses too fast to right
(2) Too much lateral movement. |

The *Gray Score Card* has been previously described as a useful device for diagnosing certain types of handwriting disabilities. The different characteristics of writing as identified by Gray are weighted in accordance with his judgment of their relative importance. The total of all assigned values may be taken as a measure of general merit in writing. However, it is probably better to use either the *Thorndike* or the *Ayres Scale* to indicate general writing ability. The *Freeman Diagnostic Chart* is a highly valuable aid to the teacher in identifying the various faults in writing and pointing out those elements which are seriously in need of improvement.

Remediation through Physical Conditions and Materials. Prominent among the physical factors affecting the pupil's handwriting is his desk. The pupil's desk should be adjusted to his height so that when he is seated normally his thigh is at right angles to the lower part of his leg and his feet are flat on the floor. In accordance with most modern methods of writing, the pupil's body, when he is writing, should face the middle of the desk squarely and bend slightly forward at the hips. Both forearms should be well up on the desk, the left holding the paper, the right wrist raised and inclined slightly to the right. It is necessary

that the pupil be taught to move the paper upward and to the left as the writing progresses. The shifting is done with the left hand, while the right arm is held in the correct position. There is some difference of opinion as to the best position of the writing arm. It is generally agreed, however, that the writing hand should be supported on the third and fourth fingers and that the wrist should not be tilted more than 45 degrees. The forearm of the right hand should be perpendicular to the line of writing. The pen should be grasped lightly and in such a way that the forefinger is below the thumb and at least one inch above the point of the pen.

Experiment and observation shows that writing is a combination of whole-arm, forearm, wrist, and finger movements. It is not possible or even desirable to eliminate entirely finger movement, even in so-called "muscular movement writing."

Remediation of Psychological Conditions. Next in importance in preparing the way for effective mastery of writing faults is the provision of desirable psychological conditions. The establishment of a desire for improvement on the part of the pupil is essential. One plan which has been proved to be quite effective involves the pupils' use of handwriting scales for the appraisal of their own writing. A copy of some good general merit scale should be conveniently placed in every classroom to encourage pupils in its use as a means of facilitating comparisons and evaluation of personal products.

Another means of motivation is the exemption from further penmanship drill of all pupils who have attained the accepted standard of speed and quality. The standard of 60 for speed and quality is the one generally accepted. Evidence seems to indicate that from 50 to 75 per cent of the pupils in the upper grades can easily reach this standard. If these pupils are exempt from further drill, the teacher is able to devote more time to those who have failed to meet the standard.

For improving the rate of writing for the inferior writers, Monroe⁸² suggests the writing of the following sentence,

⁸² W S Monroe, *Measuring the Results of Teaching*, p 227 Houghton Mifflin Co., Boston, 1918

"The quick brown fox jumps over the lazy dog," in time limits as prescribed below:

Eighth grade pupils should do this 11 times in 4 minutes									
Seventh	"	"	"	"	"	8	"	"	3
Sixth	"	"	"	"	"	6	"	"	3
Fifth	"	"	"	"	"	5	"	"	2 $\frac{3}{4}$
Fourth	"	"	"	"	"	4	"	"	2 $\frac{1}{2}$
Third	"	"	"	"	"	3	"	"	2 $\frac{1}{8}$
Second	"	"	"	"	"	2	"	"	2

This sentence provides drill on all the letters of the alphabet and should be continued in use until the pupils can write it at the speed indicated in the schedule given above. Often careful instruction in the making of different letters is needed because many children are unable to make particular letters. Pupils are greatly helped by practice upon the letters which give them trouble until the accepted standards for speed have been attained.

Handwriting drill in the past stressed the development of "movement" by means of exercises designed to develop coordination of the motor abilities needed in writing. All pupils in the group were subjected to the same exercises regardless of degree of skill or particular needs. Freeman and Dougherty³³ have prepared a number of graded drills for both "movement" and rhythm in writing designed to eliminate writing defects as revealed by the *Freeman Diagnostic Chart*. These exercises are intended for group work but may be used for individual drill.

The *Courtis-Shaw Standard Practice Tests in Handwriting* provide for complete individualization of handwriting drill. The basic principles of method underlying the *Courtis-Shaw Practice Tests* are as follows:

1. Present a definite objective goal for each child.
2. Let him try to reach it
3. Have him measure his own success or failure.
4. In event of failure, supply such assistance as he may ask for and encourage him to try again
5. In event of success, present a new and slightly more difficult objective until the ultimate goal has been reached.

The Courtis-Shaw material provides for a preliminary test to reveal the initial standing of the pupil. Those children

³³ Freeman and Dougherty, op. cit.

who meet the standard agreed upon are excused from further drill. All other children start with different practice exercises according to their needs. Two series of graded exercises are provided, one for Grades 3 to 5, and one for Grades 6 to 8. The exercises are provided with standards for both speed and quality. As children meet the standards of both rate and quality in a lesson, they go to the next, until all of the lessons are completed. Such a procedure obviously results in great variation in progress, but most certainly possesses great motivating power. The authors of these practice tests have made provision for group competition, and, what is more significant, competition with one's own records. Provision is also made for diagnosis of defects.

The *Minneapolis Self-Corrective Handwriting Charts* devised by Nystrom are also excellent devices for individualizing handwriting practice. These charts differ from the Leamer Tests and the Curtis Tests in that they do not provide for progress through graded series of exercises. They are designed to aid the pupil to discover his weaknesses and to suggest to him the drills that will enable him to overcome his defects. The underlying principles of this plan are stated as follows:

1. Individualization of instruction to meet individual needs, diagnosis of difficulties, provision of specific remedies for these difficulties.
2. Socialization of instruction through problems requiring cooperative group activity and through provision of known goals to be reached.
3. Vitalization of instruction of the child by considering handwriting not a form of penmanship drill, but the daily work in all subjects.

Two principal devices for evaluation are provided:

1. A rating scale, standardized for the Minneapolis Public Schools, similar to the Ayres Writing Scale.
2. A series of diagnostic charts. After writing has been scored on the basis of general merit, if found to be unsatisfactory, a second evaluation, diagnostic in nature, is made from a set of diagnostic charts. Included in this diagnosis are nine points

(1) Color	(5) Letter spacing
(2) Size	(6) Beginning and ending strokes
(3) Slant	(7) Word spacing
(4) Figure form	(8) Alignment
	(9) Letter form

Each pupil compares his writing with these charts, and discovers his particular defects. He then studies the reverse side of the chart for the remedial exercises which will enable him to eradicate his defects.

Leamer's *Practice Sentences in Handwriting* also provide for individual progress through a series of graded lessons very similar to the *Courtis-Shaw Practice Tests*. This system attempts to provide practice in writing letters and words that are most frequently used. The words and sentences used incorporate the words from the *Ayres Spelling List* that are most often used in life outside the school. In one arrangement, the set includes alphabet cards for each pupil, practice cards for sentences, a handwriting scale, a diagnostic chart of illegibilities in handwriting, and suggestions for follow-up work.

Handedness as a Factor in Writing. In addition to the physical and psychological conditions discussed in the preceding paragraphs, there is the very important factor of handedness in the pupil. The general considerations of method and remedial procedures in handwriting appear to assume right-handedness in the child. Yet left-handedness is common enough in the classroom to represent a significant problem to the teacher, and one worthy of some consideration here. Naturally enough the pursuit of methods of instruction and remedy suitable for the right-handed pupil results in the formation of atrocious writing habits for the left-handed pupil. Any attempts to force him to conform to common right-handed practices usually forces him to write backwards, i.e., toward the left. In order to correct for the resultant reversal of the image, the pupil frequently twists his left wrist around in such a way that the pencil or pen-point is directed toward him, with the result that he works awkwardly and under a most severe muscular maladjustment. For these and for other reasons which appear to be related to the speech and language functions, the teacher should probably not attempt to change over the left-handed pupil. It is almost certainly better to accept the tendency to left-handed dominance which is well developed by the time the child enters the first grade and merely aid him in making

the best possible adjustments and adaptations in his mastery of handwriting.

TOPICS FOR DISCUSSION

1. What are the major situations in life in which language is used?
2. Evaluate the relative demands made by life situations on the oral and written aspects of language.
3. From the standpoint of classroom emphasis, should oral language or written language receive more emphasis?
4. According to Travis and Blanton, what are the major causes of oral language disabilities?
5. Discuss the measurement of oral language abilities.
6. How is ability in written composition measured?
7. Discuss the status of analytical testing of written language abilities.
8. Discuss some remedial drill materials of value in language instruction.
9. What appears to be the most acceptable fundamental assumption upon which the spelling vocabulary suitable for elementary school instruction should be based?
10. Show how a spelling test can be made from a standard spelling scale.
11. How may a spelling test made up of socially useful words be validated for use in a classroom in which a textbook in spelling based on a vocabulary of unknown social significance is in use?
12. What range of difficulty in words would you select for the purpose of measuring a class of unusually poor spelling ability?
13. Discuss the pupil work habits which have diagnostic significance in the field of spelling.
14. Which of the objectives or outcomes of instruction in handwriting are most defensible from a social utility point of view?
15. What is the relationship between handwriting speed and quality?
16. What place have standards in the evaluation of handwriting?
17. Describe some of the methods of diagnosing handwriting ability

SELECTED REFERENCES

- Almack, John C, and Staffelbach, E. H, "Spelling Diagnosis and Remedial Teaching" *Elementary School Journal*, 34 341-50, January 1934
- Breed, Frederick S, *How to Teach Spelling*. Danville, N. Y. F. A. Owen Publishing Co, 1930
- Broom, M E, *Educational Measurements in the Elementary School*, pp. 147-82. New York McGraw-Hill Book Co, Inc, 1939.
- Brueckner, Leo J, and Melby, Ernest O., *Diagnostic and Remedial Teaching*, Chapters IX-XI. Boston Houghton Mifflin Co., 1931.
- Davis, Georgia, "Remedial Work in Spelling." *Elementary School Journal*, 27 615-26, April 1927.
- Dykema, Karl W, "On the Validity of Standardized Tests of English Usage." *School and Society*, 50.766-68, December 9, 1939.

- Foran, Thomas G., *The Psychology and Teaching of Spelling*. Washington, D. C. The Catholic Education Press, 1934.
- Freeman, Frank N., "Contributions of Research to Special Methods Handwriting" *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter VI, pp. 91-97. Bloomington, Ill. Public School Publishing Co., 1938.
- Freeman, Frank N., "Handwriting." *Encyclopedia of Educational Research*, pp. 555-61. New York: The Macmillan Co., 1941.
- Freeman, Frank N., and Dougherty, Mary L., *How to Teach Handwriting*. Boston: Houghton Mifflin Co., 1923.
- Gilliland, A. R., Jordan, R. H., and Freeman, Frank S., *Educational Measurements and the Class-Room Teacher* (Revised Edition), Chapters V, VI, VIII. New York: The Century Co., 1931.
- Greene, Harry A., "Contributions of Research to Special Methods English Usage." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter IX, pp. 115-21. Bloomington, Ill.: Public School Publishing Co., 1938.
- Greene, Harry A., "English—Language, Grammar and Composition" *Encyclopedia of Educational Research*, pp. 446-61. New York: The Macmillan Co., 1941.
- Greene, Harry A., and Betts, Emmett A., "A New Technique for the Study of Oral-Language Activities." *Elementary School Journal*, 33 753-61, June 1933.
- Hawkes, Herbert E., Lindquist, E. F., and Mann, C. R. (Editors), *The Construction and Use of Achievement Tests*, Chapter VIII. Boston: Houghton Mifflin Co., 1936.
- Horn, Ernest, *A Basic Writing Vocabulary: 10,000 Words Most Commonly Used in Writing*. University of Iowa Monographs in Education, First Series, No. 4. Iowa City: University of Iowa, 1926.
- Horn, Ernest, "Contributions of Research to Special Methods Spelling" *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter VIII, pp. 107-14. Bloomington, Ill.: Public School Publishing Co., 1938.
- Horn, Ernest, "Spelling" *Encyclopedia of Educational Research*, pp. 1166-83. New York: The Macmillan Co., 1941.
- Koos, Leonard V., "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools." *Elementary School Journal*, 18 423-46, February 1918.
- McKee, Paul, *Language in the Elementary School*. Boston: Houghton Mifflin Co., 1939.
- Madsen, I. N., *Educational Measurement in the Elementary Grades*, pp. 160-80. Yonkers-on-Hudson, N. Y.: World Book Co., 1930.
- Mort, Paul R., and Gates, Arthur I., *The Acceptable Uses of Achievement*

- Tests*, Chapter VI New York Bureau of Publications, Teachers College, Columbia University, 1932.
- Nelson, M. J., *Tests and Measurements in Elementary Education*, Chapter VII. New York The Cordon Press, 1939
- Smith, Dora V., "Diagnosis of Difficulties in English" *Educational Diagnosis* Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XIII, pp. 229-67. Bloomington, Ill Public School Publishing Co., 1935
- Smith, Henry L., and Wright, Wendell W., *Tests and Measurements*, Chapters VII-VIII. New York Silver, Burdett and Co., 1928.
- Spache, George, "Spelling Disability Correlates I—Factors Probably Causal in Spelling Disability." *Journal of Educational Research*, 34 561-86, April 1941.
- Thorndike, Edward L., *A Teacher's Word Book* (Revised Edition). New York Teachers College, Columbia University, 1931.
- Tiegs, Ernest W., *The Management of Learning in the Elementary Schools*, Chapters VII-VIII. New York Longmans, Green and Co., 1937.
- Tiegs, Ernest W., *Tests and Measurements in the Improvement of Learning*, pp. 132-37, 171-84 Boston Houghton Mifflin Co., 1939.
- Travis, Lee Edward, "Diagnosis in Speech" *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XIX, pp. 399-434. Bloomington, Ill Public School Publishing Co., 1935
- Traxler, Arthur E., *The Use of Test Results in Diagnosis and Instruction in the Tool Subjects* (Revised) Educational Records Bulletin No. 18, pp. 31-43, 62-73. New York Educational Records Bureau, January 1937.
- Webb, L. W., and Shotwell, Anna Markt, *Testing in the Elementary School*, Chapters XI-XIII. New York Farrar and Rinehart, Inc., 1939.
- Wilke, Walter H., "The Development and Application of a Scale for Measuring Diction" *Quarterly Journal of Speech*, 24 268-81, April 1938.
- Wilson, Guy M., and Hoke, Kremer J., *How to Measure* (Revised and Enlarged Edition), Chapters II-III, VI-VII. New York The Macmillan Co., 1929.

CHAPTER XVII

MEASUREMENT IN THE SOCIAL STUDIES

This chapter summarizes the following points in the improvement of instruction in the social studies.

- a. Shifting objectives in the social sciences.
- b. Measurable qualities in history, civics, and geography.
- c. Measurement of acquisitive skills, and other outcomes.
- d. Factual tests in history, civics, and geography.
- e. Problem-solving tests in the social studies.
- f. Measurement of social attitudes.
- g. Standardized social studies tests.

The social studies deal primarily with past and current problems of human relationships. The subject matter is concerned with the interactions of human beings as they associate with one another in varied political, economic, and social activities. Accordingly, the formulation of exact objectives in the social studies is difficult. In fact, there are no scientifically established objectives, such as may be found in spelling, reading, language, or arithmetic, for research techniques of determining objectives are very difficult to apply in the social studies

I. AIMS AND ORGANIZATION OF THE SOCIAL STUDIES

Objectives of the Social Studies. Despite the limitations placed upon the establishment of objectives in the social studies, many such lists have been proposed. One of the best of these lists is that given in the Report of the Commission of the American Historical Association. The following summarization of this list was prepared by Wesley:¹

- I. Information
- II. Skill in (1) using libraries and institutions, (2) using books and materials, (3) sifting evidence, (4) analysis, (5) observation, (6) writing, (7) making maps, charts, etc., (8) memorizing, (9) using the scientific method

¹ Edgar B Wesley, *Teaching the Social Studies*, pp. 170-71 D C. Heath and Co., New York, 1937.

- III. Habits of (1) neatness, (2) industry, (3) promptness, (4) accuracy, (5) cooperation, (6) economy of time and money, (7) patience.
- IV. Attitudes of (1) respect, (2) appreciation, (3) admiration, (4) faith, (5) responsibility, (6) helpfulness, (7) sympathy, (8) patriotism, (9) tolerance, (10) fairness, (11) broad-mindedness
- V. Qualities of (1) independence, (2) will power, (3) courage, (4) persistence, (5) alertness, (6) imagination, (7) initiative, (8) creativeness

The student should note that these objectives are listed as informations, skills, habits, attitudes, and qualities. The best modern thinking in the social studies results in objectives of this tangible and definite type rather than in the wordy and frequently indefinite and intangible objectives which were customarily listed until recently.

Organization of the Social Studies. The question of whether to organize the subject matter of the social studies according to the traditional subject divisions or to integrate the various fields into a unified course of study is receiving much attention from students in this field. Unified courses are based on the theory that life consists of actual problems and that the best way to prepare children to meet such problems in life is to disregard subject divisions and assemble materials from all sources possible, putting the materials together in such ways as may prove most effective in meeting the pupil's needs for problem solving. Thus, believers in the unified course would ignore geography, history, and civics as separate subjects, and would embody material from all of them in a single composite course. There is a distinct tendency toward this integrated course, especially in the elementary grades, with some approval for its adoption in the junior and senior high school grades.

While the results of comprehensive investigations show a preponderance of opinion in favor of unification of the social studies in the elementary school, curriculum-makers have not as yet evolved many such courses. In the meantime, schools generally continue to teach their subject matter as traditionally organized. Testing necessarily lags behind the development of the curriculum in this field. As the curriculum-workers progress in their work of integration, there will be a real need for standard tests in the newly organized

subject. Naturally such tests will be developed in accordance with the objectives set forth in the best available courses of study.

II. MEASURABLE QUALITIES IN HISTORY, CIVICS, AND GEOGRAPHY

Measurement of Acquisitive Skills. Efficiency in the study of the content subjects depends to a marked degree upon the pupil's mastery of reading skills of the work-study type. Instead of a few textbooks relating to a limited number of topics, the up-to-date school provides wide reading opportunities as a means of enriching the course of study. Effective reading habits are therefore essential to such content subjects as history, geography, and civics. It follows, therefore, that good teaching in the social studies, and in all of the content subjects for that matter, must provide for the improvement and refinement of the reading habits and skills required in these subjects.

Standard tests designed to measure a pupil's reading rate and comprehension in social studies content are not generally available. However, the teacher may develop his own informal tests of social science reading skill by formulating comprehension questions based upon the reading material sampled from the pupil's textbook. A careful analysis of the pupil's answers and his rate of reading should prove of value to the teacher in aiding him to improve the pupil's specific reading skills in the social sciences. The reader is referred to Chapter XV, for further suggestions in this connection.

The element of subject-matter vocabulary is basic to effective social science reading. Reading vocabulary has only recently received the attention it deserves as a factor conditioning the efficiency of reading and study. Pressey² found that much of the difficulty which pupils have in studying their textbooks was due to lack of knowledge of the more or less technical words in a subject, rather than to lack of any

² L. C. Pressey, "An Investigation of the Technical Vocabularies of the School Subjects" *Educational Research Bulletin*, 3 182-85, April 1924.

general silent reading ability. Tormey³ has also shown that vagueness in the meanings which children attach to apparently simple phrases in history content definitely limits their accomplishment in the subject. He further showed that relatively brief periods of training in the acquisition of clear-cut meanings for these terms pay big dividends in greatly improved accomplishment in the subject.

Measurement of Outcomes. The difficulty of measuring the general outcomes of the social studies is obvious. Thus far there has been apparently too little careful analysis of the several subjects into the desired knowledges, skills, habits, ideals, and attitudes to permit exacting curriculum and test construction. For most instructional purposes and for practically all testing purposes, the basic knowledges involve the retention of a few facts with accuracy, others in approximation. These facts may deal with dates, men, events, or movements. To this series of facts may perhaps be added a rough idea of their time sequence and their relational interdependence. In civics the basic knowledges include an understanding of the principles of government and their operation in the lives of men. Basic knowledges in geography include memory of facts, skill in map interpretation, technical vocabulary, the interpretation of many and varied items of information in terms of their relationships, and the ability to solve problems dealing with human relationships.

The real deficiency in existing tests in the social studies is not that they are designed primarily to measure the informational aspects of the subject, but that other abilities more important to the attainment of the larger objectives of social studies instruction have not been provided for. Unfortunately, when a standardized test is used, the particular outcomes which it measures tend to be given special attention by both teacher and pupils. As a result, important objectives other than those emphasized in the tests are likely to be neglected. The teacher must assume the responsibility for the provision of instruction on other types of outcomes. The

³ T. J. Tormey, "The Effect of Drill Upon the Specific and General Comprehension of Historical Content" Abstract in *Doctoral Theses in Education, I. University of Iowa Studies in Education*, Vol IX, No 1, pp 153-82. University of Iowa, Iowa City, 1934

measurement of such instruction will for the most part necessarily be taken care of by teacher-made tests, at least for the immediate future.

Kinds of Tests in History, Civics, and Geography. The selection of the basic facts to be taught and tested is one of the very serious problems of measurement in the social studies. The available body of facts in geography, history, and civics is large and the rapid pace of events today results in constant and great increases in the subject matter of these fields. It is not so much the need for knowledge of the array of facts as it is the determination of those likely to last in a rapidly changing world long enough to deserve special emphasis in instruction and in testing which complicates the problem. This situation has resulted in the production of relatively few tests in the social studies. In their efforts to meet the problem of which facts to teach and test, most workers in these fields have made their courses of study and their tests more and more comprehensive, hoping thereby to satisfy the ideas of all as to the basic items. Too often this has led both teacher and pupils to emphasize mere memorization of facts. As a result, these facts are too frequently mastered merely as facts, and not in order that they may give the pupil a better understanding of life and human relationships.

The majority of the tests available at present in the social science fields are of doubtful value for diagnostic purposes. Three general groups of tests in the social studies may be identified: (1) tests of facts and information, (2) tests of ability to solve social problems, (3) tests of civic, social, and economic attitudes.

Factual Tests. Tests of facts and information are by far the most numerous of the tests in the social studies. This is to be expected, for the pupil's knowledge of certain facts or items of information is quite easily discovered. Furthermore, teachers of the social studies, perhaps more than those of most other groups, have emphasized the acquisition of facts and information to the practical exclusion of other desirable general outcomes of instruction. Factual tests are of limited value for diagnostic purposes. They fail to reveal

why pupils do not know the facts if they have not been acquired. The factual tests do not aid the teacher very significantly in discovering the ability of pupils to use facts in their thinking in the social science fields.

Problem-Solving or Thought Tests. The development of the ability to utilize facts and basic principles in the attack on a novel social situation is one of the basic outcomes of teaching in the social studies. This type of problem-solving duplicates the steps in the ordinary process of thought. As in arithmetic, problem-solving in the social studies involves reading the problem to comprehend it, picking out the facts which are pertinent to the problem, choosing a method of solution, and testing the results for accuracy and probability.

It is well recognized that knowledge of the facts necessary for the solution of a problem is no guarantee that the problem will be solved. Neither can a problem be solved unless the necessary facts are available. However, availability of facts in this day of widely-available library facilities does not depend only upon a knowledge of them by their prospective user. Many of the tests for various types of problem solving abilities present the necessary facts to the pupils in the test so that the result will depend upon their abilities so to use the facts that they are able to solve the problems.

Attitudes Tests. Since our actions depend to such a large degree upon attitudes and emotional reactions, the measurement of attitudes resulting from instruction in the social sciences is as greatly needed as are tests of ability to solve problems. As a matter of fact, much attention is now being given in school to the development of desirable traits of citizenship which are so much needed in later adult life. However, up to date the measurement of such traits by attitudes tests has been largely unrealized.

The attitudes inventories available are in the main better adapted to the secondary than to the elementary school level. Furthermore, most of these attitudes tests are not devised for use in particular courses or subjects, with the exception of scientific attitudes tests for use in science courses. Even the *Thurstone Scales of Social Attitudes* and the similar generalized attitudes scales devised by Remmers and his associ-

ates are devised for attitude measurement in general rather than for use in particular courses. Perhaps a major reason why attitudes have not been more adequately measured in courses or subject fields is that even subject-matter specialists are agreed only on the broad attitudinal lines and not on the specific or functional attitudes which are desirable. The lack of agreement common on specific political, economic, national, and international issues among people who have a favorable attitude toward democracy illustrates this fact.

III. STANDARDIZED SOCIAL STUDIES TESTS

Standardized Tests in Social Studies Subjects. Most of the standardized tests which are now available for history, civics and government, and geography were published some years ago, so that it is largely in the form of a few tests for general social studies and the social studies parts of achievement test batteries that new standardized tests have appeared for this field. This may be the result of the rather slow trend toward unification of the social studies in the elementary school.

History. Standardized history tests for the elementary and junior high school grades are entirely for American history, in order to conform to the course offerings below the high school level. The major emphasis of most tests is upon factual knowledges, although some of the tests satisfactorily measure some of the more complex and significant results of instruction requiring various applications and interpretations of factual data. Too much emphasis doubtless appears in many tests upon specific facts rather than upon approximate knowledges, particularly in connection with exact dates.

Civics and Government. Standardized tests in the field of civics and government are limited in number. In general, measurement here is as satisfactory as could be expected under the changing conditions now existing in the social studies. However, there is need for tests which attack citizenship problems in a more positive and realistic manner than do most of the standardized tests in civics now available.

Geography. Many tests are available in geography, but most of these are of the formal factual type. Few of the tests take into account the problem-solving aspects of social studies instruction. The majority of standardized tests in geography attack the subject as a study of places and their characteristics, whereas the modern approach to the study of geography has come largely to be founded upon the manner in which geographical factors influence human beings and the societies which they establish.

Standardized Tests in General Social Studies. Makers of standardized tests have begun to develop tests in the social studies at the junior high school and even the elementary school level to meet the needs of schools which may be offering the unified type of social studies course discussed in a preceding section of this chapter. Tests of this type are not uncommon for the high school, but practically all of the social studies tests for the elementary and junior high school grades have been for particular courses until the last few years. These tests include material from history, from civics and government, and from geography, but subject matter lines are broken down.

Tests which measure broadly over the social studies must almost of necessity avoid some of the weaknesses of tests in particular subjects because of their lack of concern for divisions within the field. Furthermore, the few tests of this type are relatively new, and consequently have the advantage of being constructed with regard for recent thinking and experimentation with tests. Factual knowledges are less stressed and greater emphasis is placed upon relationships, applications, interpretations, and other reasoned uses of facts than is true on the average of standardized tests for particular courses at the elementary level.

Items of these tests are similar in form, but not in content, to the multiple-choice and matching forms presented in the following section of this chapter. It is impossible because of space limitations to present sample items here which would serve to illustrate these tests adequately, but listings of the abilities they test will serve to indicate their scope and emphases.

The *Cooperative Social Studies Test for Grades 7, 8, and 9* has three parts: (1) facts, skills, and applications, (2) terms and concepts, and (3) comprehension and interpretation. The *National Achievement Social Studies Test* for grades 4 to 6 deals with: (1) human relations, (2) life situations, (3) social problems, (4) products and peoples, and (5) the meaning of events. Major outcomes tested by the *Wrightstone Test of Critical Thinking in the Social Studies* for Grades 4 to 6 are: (1) obtaining facts, (2) drawing conclusions, and (3) applying general facts. It is significant that nowhere in the lists of parts for these three tests is there any indication of division along the subject lines by which the social studies field is usually divided.

Standardized Test Methods. The practice of presenting illustrative types of objective items to familiarize the student with representative measurement techniques is followed in this chapter. A desirable degree of knowledge on the part of the student concerning specific standardized tests could not be assured in the brief treatment possible here through the discussion of particular standardized tests accompanied by occasional illustrations. The student can gain such knowledge of particular tests only by examining them critically or even administering them to groups of pupils under standard conditions.

A few sample items representative of test item techniques used by makers of standardized history, civics, geography, and general social studies tests are given in this section. These should serve the double purpose of acquainting students and teachers with standardized testing techniques and of suggesting to them types of test items and exercises they can construct for their own informal objective tests.

Simple Recall and Sentence-Completion Items. The simple recall and sentence completion items are not widely used in standardized social studies tests. Only three samples of these forms are presented, for differences among various recall items exist mainly in minor details. The first sample is of the basic simple recall form, the second is of simple recall items based on a diagram, and the third is from a sentence completion exercise.

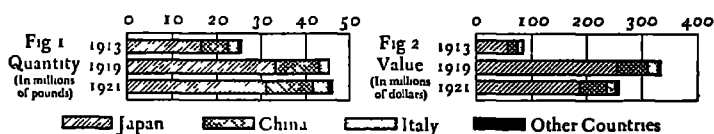
Sample A.⁴

- 41 The successor of McKinley to the presidency was named (41) _____
- 42 The battle cry of the Texan Army was "Remember the (42) _____
- 43 The ship in which Henry Hudson first sailed up the Hudson River was called the (43) _____

Sample B.⁵

DIRECTIONS. Read the questions below. Find in the graphs the answer to each question and write the answer in the parentheses after the question.

IMPORTS OF RAW SILK FOR THE UNITED STATES



26. What country is third in exporting raw silk to the United States? () 26
27. In what year did the United States import the least silk? () 27
28. What year ranks second in the value of raw silk imported? () 28

Sample C.⁶

The ship, _____, built in England, destroyed
40
many Northern merchant ships. Later the Kearsarge destroyed it.

The battle between the _____ and the _____
41 42
proved that ironclads were to take the place of wooden war vessels

⁴ M H DeGraff, G M Ruch, and H A Greene, *Iowa General Information Test in American History*, Grades 7-12 Published by Bureau of Educational Research and Service, University of Iowa, 1927

⁵ N Theresa Wiedefeld and E Curt Walther, *Wiedefeld-Walther Geography Test*, Grades 4-8 Published by World Book Co, 1931

⁶ Lena A Ely and Edith King, *Ely-King Tests in American History*, Junior High School Published by Southern California School Book Depository, 1927.

Alternate-Response Items. The majority of alternate-response items in elementary social studies tests are of the true-false or yes-no variety, although other forms occasionally occur. The illustrations are of a "plus-zero" (true-false) item type, a special adaptation of the alternate-response item, and a modified true-false form which asks pupil to indicate correct answers by a "C" and to write in the words which make the false items incorrect.

Sample D.⁷

46. The "underground railroad" aided in the enforcement of the Fugitive Slave Law () 46
47. The majority of immigrants who came to the United States before 1870 came from Germany and the British Isles. () 47
48. The opening of big Western farms decreased the need for agricultural machinery () 48

Sample E.⁸

III. Place a cross before the event which came first in each of the following groups:

- 21 () Beginning of Mexican War or
() Annexation of Texas
22 () Admission of California as a state or
() Discovery of gold in California

Sample F.⁹

- A The capital of Illinois is Chicago.
B. The interior plateau of South Africa _____
is called the veldt.
41. The shape of the earth is round.

Multiple-Choice Items. The multiple-choice item appears to be the most popular for testing purposes in the elementary social studies subjects. The illustrations given

⁷ Harry J. Carman, Thomas N. Barrows, and Ben D. Wood, *Junior American History Test*, Junior High School. Published by World Book Co., 1929.

⁸ Ely and King, *op cit*.

⁹ Alice McGill, *Every Pupil Test of Geography*, Grade 7. Published by State Department of Education, Ohio, April 1940.

below are of the simple item form, of the common form based on a paragraph of reading material in civics, and of a multiple-response item.

Sample G.¹⁰

13. Marco Polo traveled in—

- 1 Mexico, 2 Canada, 3 China, 4 South America,
5 Japan ()

14. The great Spanish Armada was defeated by—

- 1 France, 2 Portugal, 3 Holland, 4 the Turks,
5 England ()

Sample H.¹¹

1. Suppose that a sensational hold-up has been staged by three masked bandits in which \$10,000 worth of stocks, bonds, and valuable papers have been stolen from a bank in the heart of the city. In the chase that ensued after the discovery of the robbery, one policeman was fatally wounded and another seriously injured by bullets from the guns of the bandits. Suppose that you know that the bandits have taken refuge in a house in your neighborhood. As a citizen, why should you give this information to the police?

1. Your own life is in danger.
2. Your home may be robbed of valuables.
3. You have always feared bandits.
4. Criminals against society should be restrained.
5. You may receive a large reward.

Sample I.¹²

7. The main causes of the westward movement (before 1860) were:

- The discovery of gold in California _____
 The ease of living in the West _____
 The liberal land policy followed by the national government _____
 The adventurous spirit of the American people _____
 The completion of the transcontinental railroad _____

Matching Exercises. Matching exercises appear to be second to multiple-choice item forms in popularity for the

¹⁰ E C Denny and M J Nelson, *Denny-Nelson American History Test*, Grades 7 and 8 Published by World Book Co, 1928

¹¹ Arold W Brown and Clifford Woody, *Brown-Woody Civics Test*, Grades 7-12 Published by World Book Co, 1926

¹² A S Barr and C J Daggett, *Information Tests in American History*, Grades 7-12 Published by Educational Test Bureau

testing of achievement in elementary social studies courses. Not only are balanced matching exercises widely used, but the unbalanced matching and exercises based on the multiple use of one category of items and on graphs or maps are also common. An illustration is given below for each of these four types.

Sample J (Exercise shown only in part).¹³

Champlain (1)	36. Pioneered in Kentucky	. . (
James Wolfe (2)	37. Captured Fort Ticonderoga	. . (
Ethan Allen (3)	38. Helped to settle Jamestown	(
John Winthrop (4)	39. Famous missionary to the Indians	(

Sample K.¹⁴

COLUMN 1	COLUMN 2
(STATEMENTS OF EFFECT)	(STATEMENTS OF CAUSES)
SAMPLE.	
Fishing in Norway (3)	1. high, snow-capped mountains
1. Wandering life of the Eskimos ()	2. people depending on wild animals for food
2. Transportation by camel ()	3. poor, rocky, forest-covered soil near the sea
3. Houses of wood and bark with steep roofs	4. cool, damp climate, much low, wet ground
of leaves ()	5. hot, dry climate, much soft, loose sand.
4. People wearing wooden shoes when working ()	6. hot, rainy climate, many raffia-palm trees
	7. broad, level plains, rich soil, moderate summer rain.

Sample L.¹⁵

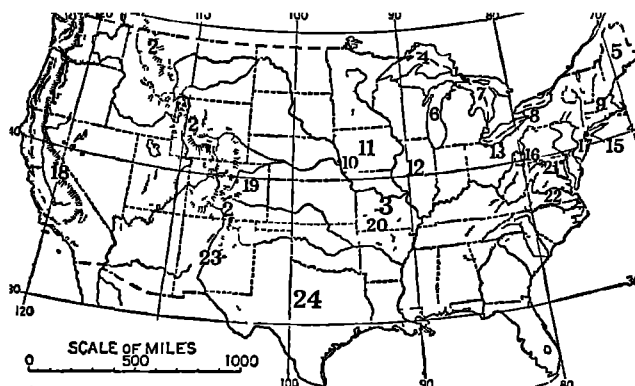
II. Place a "B" before the name of each country in the following list that sent many immigrants to the United States between 1860 and 1890. Place an "A" before those that furnished most of the immigrants after 1890:

- | | |
|-----------------|-------------------------|
| 11. () Russia | 14. () Italy |
| 12. () Ireland | 15. () Germany |
| 13. () England | 16. () Austria-Hungary |

¹³ Denny and Nelson, *op cit*

¹⁴ Wiedefeld and Walther, *op cit*

¹⁵ Ely and King, *op cit*

Sample M.¹⁸

- | | | | |
|-----------------------------|-----|-----------------------------|-----|
| 13. New York City | () | 20. Lake Michigan | () |
| 14. Massachusetts | () | 21. Texas | () |
| 15. Baltimore | () | 22. Pittsburgh | () |
| 16. Cleveland | () | 23. Sierra Nevada | () |
| 17. Ozark Plateau | () | 24. Omaha | () |
| 18. Missouri | () | 25. New Jersey | () |
| 19. Denver | () | 26. Buffalo | () |
| | | 27. Richmond | () |

IV. INFORMAL OBJECTIVE TESTS IN THE SOCIAL STUDIES

Teacher-Made Objective Tests in the Social Studies.

More attention has been given to informal objective testing methods for the social studies of the high school level than of the elementary school level in the educational literature, except possibly for geography. This does not by any means imply that informal objective tests are used only or even mainly at the higher level. It does suggest, however, that few new techniques or modifications of old techniques have been devised for the social studies below the high school. Such a situation is not surprising, in view of the fact that the small number of new standardized tests in this field are included in achievement test batteries or are for the general social studies.

Two means of evaluating instructional outcomes of the social studies informally are open to the teacher: (1) the

¹⁸ Wiedefeld and Walther, op. cit.

construction of informal objective tests, and (2) the use of other evaluative procedures. Illustrations and discussions of item types in the preceding section of this chapter should aid the teacher in constructing objective classroom tests. The program of evaluation quoted below deals largely with devices of a non-test nature.

An Evaluation Program. A comprehensive program for the evaluation of the instructional outcomes of the social studies is given below for its value in suggesting a variety of suitable measurement techniques to supplement paper-and-pencil tests. Wesley points out that although most of the suggested techniques are objective, materials for all of them are not in existence. A challenge is thereby presented to the alert teacher to devise his own evaluation instruments in such cases.

A PROGRAM OF EVALUATION¹⁷

OUTCOMES	EVALUATION TECHNIQUES
1. Concepts	Objective tests which involve at least two meanings of each significant word, tabulations of concepts used by pupils, orally and in writing
2. Study skills	Tests, completion exercises in map reading, problems in making graphs, exercises in interpreting cartoons, graphs, and tables, check-lists of pupil procedures in library and study halls.
3. Finding materials	Skills tests; check-lists for guided observation of pupils as they work, time test of skill in using index, contents, title page, card catalogue, encyclopedia, etc.
4. Information	Objective tests, class marks.
5. Reading activities	Library circulation records, records of articles and books read (cautiously compiled).
6. Interpretative reading	Tests in reading social studies materials, multiple-choice test containing elements of an outline or summary of material known to the pupil, reconstruction exercises, evaluation by the teacher of the rapid reading of material unfamiliar to the pupil.

¹⁷ Edgar B Wesley, *Teaching the Social Studies*, pp 595-96. D. C Heath and Co., New York, 1937

7. Interpreting data Tests of the relevancy of data to particular problems, of the relevancy of statements to a conclusion, exercises in grouping related sets of data, lists of data necessary to solve an assigned problem.
8. Critical attitude Tests involving the evaluation of the reliability of various sources, involving the matching of various types of persons with the fields of their probable competence, involving degrees of probable truth among various witnesses, lists of articles purchased, shows attended, and books read, with alleged reasons, tests for superstitions, a correlation of attitudes with information on the same selected topics, tests on the relevancy of various statements toward the support of a generalization or declaration.
9. Interests Actual choice of books from a varied assortment; observations of those portions of a newspaper which are being read after two minutes, observations of those subjects of magazine articles being read after five minutes, the content of pupil conversations, choice of projects and problems, games played, questionnaires, shows attended, record of hobbies, radio programs heard.
10. Cooperation Check-lists of instances of voluntary cooperation, check-lists with graded levels for indicating the quality of cooperation, lists of achievements which are the result of joint enterprises, the number and efficacy of typical student-managed organizations, check-lists of observance of courteous demeanor, tests of attitude toward cooperation.
11. Suspended judgment A test consisting of sets of statements followed by conclusions, some of which are warranted and others which are unwarranted, tests to measure the change of opinions after hearing a speech, seeing a show, reading a book, tests to see if pupils will refrain from forming judgments on insufficient bases.
12. Toleration Tests on racial and religious toleration, a check-list of instances of favorable and unfavorable treatment of minorities, such as foreigners, Negroes, etc., in the school.

V. CORRECTIVE WORK IN THE SOCIAL STUDIES

Diagnosis and Remedy in the Social Studies. Diagnosis in the social studies is difficult because: (1) the knowledges, skills, and informations which pupils should learn are not too clearly identified, and (2) if the facts to be learned were known accurately it would still be impossible to determine whether the pupil functioned in his social relationships in a desirable manner because of his possession of the informational elements revealed by a test. Diagnosis and remedy are often needed in those skills which are basic to successful work in the social studies. Instruction in these subjects requires much reading of the work-study type. Therefore, pupils, in order to achieve at acceptable levels, must possess many of the following work-study reading skills:

1. Knowledge of technical vocabularies employed in the social studies.
2. Reading comprehension with respect to adequate interpretation of social science content.
3. Ability to locate material readily—use of the index, library files, table of contents, maps, charts, etc.
4. Ability to outline
5. Ability to summarize.

These skills are discussed in Chapter XV, along with other ways and means for corrective work in these important acquisitive skills, so they are not taken up here.

TOPICS FOR DISCUSSION

1. Define the field of the social studies in such a way as to clarify the objectives adequately for testing purposes
2. Discuss the pros and cons of a unified social studies curriculum as contrasted with the traditional organization of the content by subjects.
3. To what extent do you believe the social studies teacher should emphasize the acquisitive skills lying back of achievement in the subject?
4. What procedures in the testing of problem-solving in arithmetic might be applied in the social sciences?
5. State four of the general outcomes of instruction in the social studies.
6. What are the three main types of social studies tests as specified in this chapter?
7. In your judgment what is the relation of fact to problem solving in the social studies?

8. What are the chief weaknesses in the problem-solving and the attitudes tests?
9. Discuss the use of various objective test item forms in standardized social studies tests.
10. Comment on some of the evaluative techniques of a non-test nature suggested by Wesley for use in the social studies.

SELECTED REFERENCES

- Anderson, Howard R, "Testing in the Social Studies." *Education*, 58: 545-49, May 1938
- Branom, Mendel E, *The Measurement of Achievement in Geography*. New York The Macmillan Co, 1925.
- Broom, M E, *Educational Measurements in the Elementary School*, Chapter IX New York McGraw-Hill Book Co, Inc, 1939.
- Brueckner, Leo J, and Melby, Ernest O, *Diagnostic and Remedial Teaching*, Chapter XII Boston Houghton Mifflin Co, 1931.
- Cain, Maud, "A Study of Thirteen Standard Geography Tests." *Journal of Geography*, 34 252-56, September 1935.
- Casto, E Ray, "Spices A Teaching Test" *Journal of Geography*, 38 371-72, December 1939.
- Chassell, Clara F, and Chassell, Ella B, "A Test and Teaching Device in Citizenship for Use with Junior High School Pupils" *Educational Administration and Supervision*, 10 7-29, January 1924.
- Commission on the Social Studies, American Historical Association, *Conclusions and Recommendations of the Commission*, Chapter VI. New York Charles Scribner's Sons, 1934
- Gilliland, A R, Jordan, R H, and Freeman, Frank S, *Educational Measurements and the Class-Room Teacher* (Revised Edition), Chapters X-XI New York The Century Co, 1931.
- Grim, Paul R, "A Technique for the Measurement of Attitudes in the Social Studies." *Educational Research Bulletin*, 15 95-104, April 15, 1936.
- Hamalainen, Arthur E., "Evaluation in the Social Studies." *Social Studies*, 28 250-52, October 1937.
- Hawkes, Herbert E, Lindquist, E F, and Mann, C R (Editors), *The Construction and Use of Achievement Tests*, Chapter IV. Boston: Houghton Mifflin Co, 1936.
- Kelley, Truman L, "The Objective Measurement of the Outcomes of the Social Studies." *Historical Outlook*, 21 66-72, February 1930.
- Kelley, Truman L, and Krey, August C, *Tests and Measurements in the Social Sciences*. Report of the Commission on the Social Studies, American Historical Association, Part IV. New York Charles Scribner's Sons, 1934.
- Lancaster, F, "Measuring Results in Geography." *Journal of Geography*, 30 342-45, November 1931.
- Lange, Stella R, "A Geography Test." *Journal of Geography*, 34: 40-41, January 1935.

- McCallister, James M, "Reading Difficulties in Studying Content Subjects." *Elementary School Journal*, 31 191-201, November 1930
- Madsen, I N, *Educational Measurement in the Elementary Grades*, pp. 181-88. Yonkers-on-Hudson, N. Y. World Book Co., 1930
- Michell, Elene, *Teaching Values in New-Type History Tests*. Yonkers-on-Hudson, N Y World Book Co., 1930.
- Miller, G S, "Testing Map Reading Ability." *Journal of Geography*, 30 38-42, January 1931.
- Mort, Paul R, and Gates, Arthur I., *The Acceptable Uses of Achievement Tests*, Chapter VIII. New York Bureau of Publications, Teachers College, Columbia University, 1932.
- Murra, Wilbur F, Wesley, Edgar B, and Zink, Norah E, "Social Studies." *Encyclopedia of Educational Research*, pp. 1130-56 New York The Macmillan Co., 1941.
- Neave, E M, "Geography Test, How Teachers May Construct Objective Tests." *Grade Teacher*, 50.184-85; November 1932.
- Nelson, M. J, *Tests and Measurements in Elementary Education*, Chapter VII New York The Cordon Co, 1939.
- Orata, Pedro T, "Evaluation in the Field of Social Science." *Educational Method*, 16 121-37, December 1936
- Price, Roy A, "Tests in the Social Studies" *Social Studies*, 26 23-29; January 1935.
- Ruch, G M, et al, *Objective Examination Methods in the Social Studies*. Chicago Scott, Foresman and Co, 1926.
- Smith, Henry L, and Wright, Wendell W, *Tests and Measurements*, Chapters X-XI. New York Silver, Burdett and Co, 1928.
- Tiegs, Ernest W, *The Management of Learning in the Elementary Schools*, Chapter X. New York Longmans, Green and Co, 1937
- Tyler, Ralph W, "Improving Test Materials in the Social Studies." *Educational Research Bulletin*, 11 373-79, November 9, 1932.
- Webb, L W, and Shotwell, Anna Markt, *Testing in the Elementary School*, Chapters XIV-XV. New York Farrar and Rinehart, Inc., 1939
- Wesley, Edgar B., "Diagnosis in the Social Studies." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XV, pp. 303-30. Bloomington, Ill.. Public School Publishing Co, 1935
- Wilson, Guy M. and Hoke, Kremer J., *How to Measure* (Revised and Enlarged Edition), Chapters XII-XIII. New York The Macmillan Co, 1929
- Wilson, Howard E, and Murra, Wilbur F., "Contributions of Research to Special Methods The Social Studies" *The Scientific Movement in Education*, Part II, Chapter XII, pp 147-60. Bloomington, Ill Public School Publishing Co, 1938
- Wrightstone, J. Wayne, "Measuring Some Major Objectives of the Social Studies" *School Review*, 43 771-79, December 1935.
- Wrightstone, J Wayne, "Recent Trends in Social-Studies Tests." *Social Education*, 1 246-50, April 1937.

CHAPTER XVIII

MEASUREMENT AND REMEDIATION IN THE ELEMENTARY SCIENCES

This chapter discusses the following points involved in the measurement and improvement of instruction in the elementary sciences

- a.* Aims of the elementary sciences.
- b.* Outcomes of the elementary sciences.
- c.* Limitations of measurement in the sciences.
- d.* Standardized tests in the elementary sciences.
- e.* Testing methods in the elementary sciences.
- f.* Informal objective testing of elementary science outcomes.

Introduction. This chapter supplements the preceding chapter by furnishing a further discussion of the problems of measurement in the content subjects. It deals with the elementary science fields of nature study, hygiene, and general science.

Adequate science instruction in the elementary school should be expected as a matter of course in an age of science such as the present. It is rather surprising to note, therefore, that progress in the selection and organization of science content and improvement in teaching and testing methods and materials in recent years have been slight, in spite of the great practical value of science and its natural appeal to the curiosity and interests of children. This lack of progress is particularly noticeable in the elementary school. Science as an elementary school subject has made but few contributions of experience or enrichment to the progress of education. Apparently most of our adult population have learned to make their adjustments in this scientific age through experiences they have had outside of the elementary school.

I. SCOPE OF THE ELEMENTARY SCIENCES

Aims of Elementary Science. The statement of the aims of elementary school and junior high school science

teaching given in Table XVII is adapted from Noll's compilation of the aims listed in one hundred thirty different sources.¹ Such a wide representation of opinions of science specialists should furnish an excellent indication of the outcomes toward which science teachers of the elementary school and junior high school are probably working.

It appears from the data of Table XVII that outcomes of the knowledge, habit, appreciative, interest, ability, and attitudinal types occur in both lists in that order of importance. The major differences are that knowledges are emphasized in the junior high school and habits are stressed in the elementary school, but the knowledge and habit objectives constitute more than half of the emphasis in importance at each level. Noll comments upon the surprisingly small amount of emphasis placed upon the development of attitudes and health habits.²

General Outcomes of Elementary Science. The elementary sciences must be viewed from two rather specific points of view—for their immediate educational values for children of the elementary school level, and for the background of preparation they afford for the later more intensive and specialized study of the sciences. Educational values of real significance will be attained if pupils, as a result of such instruction, acquire (1) the ability to use the scientific findings which apply in their experiences, (2) the ability to interpret natural phenomena in their environments, and (3) an appreciation of scientific attitude through understanding of and ability to use some of the methods of study which have been employed by scientists.³

The question of organization of the subject matter arises here as in the social studies. Should the sciences follow the traditional subject divisions or should they be integrated to produce a unified course of study? The tendency in the most progressive schools is toward unification. On the whole, this movement has met with more general approval in

¹ Victor H. Noll, *The Teaching of Science in Elementary and Secondary Schools*, p. 9. Longmans, Green and Co., New York, 1939.

² *Ibid.* p. 10.

³ S. Ralph Powers, "The Plan of the Public Schools and the Program of Science Teaching," *A Program for Science Teaching*. Thirty-First Yearbook of the National Society for the Study of Education, Part I, Chapter I, p. 10. Public School Publishing Co., Bloomington, Ill., 1932.

TABLE XVII

PERCENTAGE OF MENTION OF VARIOUS AIMS OF SCIENCE
TEACHING FOUND IN 130 SOURCES (ADAPTED FROM NOLL)

Aims of Science Teaching	Elementary School		Junior High School	
<i>Knowledges</i>	27 6		38 6	
1 Knowledge of the principles and applications of science		13 1		15 6
2 Knowledge leading to an understanding of the nature and organization of the environment		13 2		14 3
3 Exploration to acquaint the pupil with science and to help him to orient himself with respect to the different sciences		1 3		5 2
4 Preparation for further work in science and for college entrance		0 0		3 5
<i>Habits</i>	23 6		17 3	
5 Desirable habits of work and study		18 3		12 1
6 Habits of healthful living		5 3		5 2
<i>Abilities</i>	9 2		9 5	
7 Ability to use the scientific method		6 6		6 9
8 Ability to do useful tasks		2 6		2 6
<i>Scientific Attitude</i>	7 8		5 2	
9 Scientific attitude		7 8		5 2
<i>Interests</i>	10 5		9 6	
10 Interest in science		1 3		4 8
11 Interest in environment		9 2		4 9
<i>Appreciations</i>	15 8		14 7	
12 Appreciation of the beauties of nature and of the commonplace		14 5		10 4
13 Appreciation of the work of scientists		1 3		4 3

the elementary grades than it has at the secondary school or college levels.

Regardless of the desirability of developing an ideally integrated course of study, most of the elementary school science now taught is presented in the form of the separate units, as nature study, physiology, and general science. The scope of each of these is discussed here without reference to their possible integration.

Nature Study. The direct needs of life to which nature study contributes are of three kinds—economic, hygienic, and appreciative. Knowledge concerning how plants and animals serve our needs, involving soils, climatic conditions and effects, tillage, control of pests, plant and animal foods, and means of preserving plant and animal products, is important. Much of this knowledge is biological. There is also much need for knowledge of the physical sciences, such as the simpler operations and principles of physics and chemistry in connection with food, clothing, shelter, transportation, and other everyday problems.

The appreciative needs of nature study are of two kinds—æsthetic and intellectual. The æsthetic needs grow out of interests in the beauties of nature. The revelations of beauty in plant and animal forms, in land and water formations, and in earth and sky by day and night furnish much enjoyment. Furthermore, the cultivation of these interests affords purpose to many recreational activities. Intellectual needs resulting from the natural curiosity of children in how and why the forces of nature operate as they do are satisfied by the study of nature. By cultivation, this curiosity may be developed into a permanent interest in nature and science.

Physiology and Hygiene. The proper care of the body requires some knowledge of the structure and use of its parts and the development of proper habits in caring for these parts. The general structure of the teeth, the skin, the nails and hair, the eyes, the ears, the nose, the throat, and the mouth should be known for the contribution of this knowledge toward keeping them all in a healthful condition. In connection with the uses of food, clothing, shelter, and recreational activities, a general knowledge of the digestive organs, lungs, circulatory system, organs of excretion and sex, and the nervous system is useful in keeping these organs healthfully at work. This body of knowledge and the health habits developed with it constitute the hygienic aspects of nature study. As health knowledge tests will be considered in Chapter XX of this volume, physiology and hygiene tests are treated here only to the extent to which they enter into general tests for the elementary sciences.

General Science. Most intelligent adjustments, as distinguished from those which are purely accidental, impulsive, or habitual, are dependent upon scientific procedures. Everyone is called upon to make such responses in connection with his house, his neighborhood, his vocation, his civic duties, and his leisure. He is frequently confronted with a need for some special knowledge of health control, mechanics, chemistry, physics, biology, plant and animal life, etc. Almost every hour of the day, the individual is in the midst of the influence of mechanical and scientific appliances. For their operation, maintenance, adjustment, and repair, and as a protection from their dangers, he needs information and first-hand experience of the type obtained in general science.

II. LIMITATIONS OF MEASUREMENT IN THE SCIENCES

Difficulties in Constructing Science Tests. The construction of science tests should be relatively simple, since the content of science is quite tangible. However, difficulties of a degree no less marked than in the other content subjects are encountered. There is the same lack of agreement as to the content of the course of study and its organization as is found in the social sciences. Controversies with respect to the importance of facts as contrasted with emphasis upon relationships and problem solving are still somewhat in evidence with respect to science teaching, although scientists have increasingly of late years given attention to the more intangible outcomes of instruction. There is very little objective evidence as to what particular skills and principles, or what elements and safeguards to scientific thinking, are of most importance or can best be imparted in the elementary school sciences. The average science course apparently attempts to accomplish little more than to give a knowledge of the names of a few of the common animals, plants, and physical objects, and an acquaintance with a few of the simpler natural phenomena, without any very definite purpose appearing to justify the accumulation of such information.

Real evidences of accomplishment in the sciences are to be found in the development and the direction of pupils'

interests, attitudes, appreciations, skills, habits, ideals, and actions in these fields. The ideal way to determine the changes which are effected in the pupil as a result of studying a unit in science would be to measure the increment of desirable activities which he can and does perform as a result of this study. However, only a few attempts to devise tests for such a purpose have so far been made.

Measurable Outcomes of Science. Four major types of measurable qualities are designated in the sciences.

Facts. Most tests in science tend to over-emphasize information and knowledge as the goal of study. It is too often assumed that knowledge is a positive index of satisfactory modes of adjustment. This assumption, of course, is only partially defensible. Merely to know is no assurance of subsequent proper reaction. But, in so far as knowledge is essential to adjustment, its proper worth should not be discounted. Accordingly, measurement of the pupil's knowledge of scientific facts is to that extent valid and defensible.

Relationships. Facts in science are the vehicles for thought. The understanding of the relationships of facts and of generalized ideas is deemed most important. It is these generalized ideas which pupils should attain in their study of science. Tests should, therefore, so far as possible, measure the relational aspects of science, and do succeed in this aim to a reasonable degree.

Problem Solving. Problem tests in science call for the application of knowledge, and may demand one or more types of scientific thinking. Similarly, test items which involve the interpretation of new situations demand more than mere recall and, thus, are measures of ability to use scientific knowledge or judgment. Such test items should find a more extensive place in testing procedures than they have thus far been given.

Attitudes and Interests. Some progress has been made in the measurement of the attitudes and interests of pupils with respect to science material and phenomena. The task is a difficult one, because as yet the interests and the attitudes deemed desirable have not been defined very clearly. Furthermore, it is not too clear how the pupil should be tested,

or what should be the content of the test which will reveal the pupil's possession of the desired interests and attitudes in a dynamic sense. Some significant attempts have been made in the measurement of scientific attitudes, however.

III. STANDARDIZED TESTS IN ELEMENTARY SCIENCE

Standardized Tests in Course Areas. The number and variety of standardized elementary science tests is not great. Tests for the intermediate grades are found mainly as parts of achievement test batteries, and these parts are seldom available in separate booklets. Only one nature study test is known to the writers, although the *Modern School Achievement Test* includes some nature study materials in its elementary science section. No physiology and hygiene tests are known except the part devoted to that subject in the *New Stanford Achievement Test*, but the most recent edition, the *Stanford Achievement Test*, includes physiology and hygiene materials in the elementary science section.

Standardized tests in nature study and probably in physiology and hygiene are dependent upon (1) a more universal agreement as to their aims and purposes, (2) more representative criteria for course of study content, and (3) a more definite identification of their minimum essentials. A trend seems apparent, however, toward the merging of both of these courses with the somewhat broader course in elementary science.

Significant development of a unified course in science continues in Grades 7, 8, and 9. In the senior high school grades, the general science course has not been widely accepted, with the result that the separate subject-matter courses of biology, botany, chemistry, and physics are still taught in most high schools. Tests are available for the general science course typically given in the ninth grade. Test batteries recently published have general or natural science sections for the junior high school grades and at least two of them provide separate booklet editions for the natural sciences.

In the attempt to determine the adequacy of tests in cer-

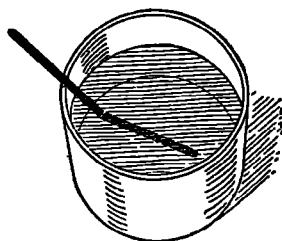
tain science fields, Diamond carefully analyzed the content of sixteen tests, most of which were standardized—eight each in general science and biology.⁴ Most of the tests were published between 1924 and 1930. He found errors of various types in more than ten percent of the more than three thousand objective items he analyzed. He classified the item weaknesses he encountered as resulting from: (1) false generalization, (2) failure to keep up with scientific progress, (3) mistaking theory for proven fact, (4) lack of scientific classification, (5) lack of scientific definition, (6) errors of tradition, (7) ambiguity, (8) spelling and typographical errors, and (9) lack of difficulty in test items. The fact that errors of these two types—in subject matter and in the mechanics of item construction—are so common in science tests, some of which are in wide use today, illustrates something of the problem involved in the construction of adequate science tests.

The *Van Wageningen General Science Reading Scales* for grades seven to twelve are designed principally to measure ability to read science material of varying degrees of difficulty with comprehension. The *Glenn-Gruenberg Instructional Tests in General Science* are designed for diagnostic and inventory purposes, and are also useful for individualized drill in the classroom. A scale of use in evaluating answers to thought questions is the *Odell Scales for Rating Pupils' Answers to Nine Types of Thought Questions in General Science*.

Standardized Test Methods. Sample items illustrative of the manner in which various objective item forms are used in elementary science testing are presented here. The student should utilize the sample items together with the bibliography at the end of the chapter for information concerning standardized tests as well as for suggestions on types of informal objective items suitable for use in the elementary sciences.

Simple Recall Items. The following samples show the manner in which a simple recall item can be used with pictorial representation.

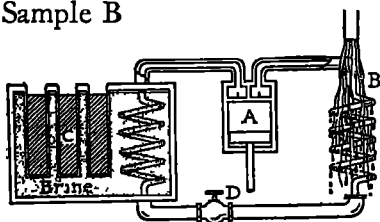
⁴ Leon N. Diamond, "Testing the Test-Maker" *School Science and Mathematics*, 32:490-502, May 1932.

Sample A.⁵

- a* This diagram illustrates the facts of the refraction of light.
- b* The amount of apparent bending of the stick depends upon the density of the liquid.

Completion Items. Simple recall and completion items differ only slightly in form and not at all in the nature of the pupil's response. The first two items below are of the simple recall or sentence completion type and the third is of completion form.

Sample B



In this drawing of an artificial ice plant

- a* The freezing vats are located at the top of the plant.
- b* The condensing pump is located at the bottom of the plant.
- c* The principle involved in the manufacture of artificial ice is that the liquid turns into a solid when the pressure is removed and, in so doing, it takes up heat from the brine, which in turn lowers the temperature of the brine.

is that the liquid turns into a solid when the pressure is removed and, in so doing, it takes up heat from the brine, which in turn lowers the temperature of the brine in the freezing vats.

True-False Items. The following samples of true-false items illustrate one of the few applications of this item type in elementary science tests.

Sample C.⁷

1. The crow is a useful bird to the farmer ()
2. Toads eat insects ()
3. Woodchucks live in the ground ()
4. Some plants grow in the water ()
5. Butterflies fly at night ()

⁵ Giles M. Ruch and Herbert E. Popenoe, *Ruch-Popenoe General Science Test*. Published by World Book Co., 1923.

⁶ Ibid.

⁷ T. L. Torgerson and Glenn A. Sealy, *Public School Achievement Tests, Nature Study*, Grades 4-8. Published by Public School Publishing Co., 1931.

Multiple-Choice Items. By far the most popular item form in elementary science tests, the multiple-choice type of item, is used in several different adaptations. Samples D to F show sample items of the common type, items based on a diagram, and items based on a passage to be read. A more comprehensive illustration of multiple-choice items from the *Unit Scales of Attainment, Elementary Science*, appears on page 13.

Sample D.⁸

- 1 Winds are — 1 rain clouds 2 moving air 3 storm clouds 4
 2 The buzz of a fly is made by its — 4 wings 5 feelers 6 legs 1
 3 The heart pumps — 7 water 8 air 9 blood 1
 4 A bird that builds its nest on the ground is the — 1 meadow lark 2 blue jay 3 oriole
 5 Nicotine is a — 4 drink 5 drug 6 food 1

Sample E.⁹

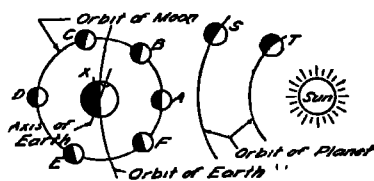


FIG. 3

33. Fig 3 When are the bodies represented at S and T visible?
- 1) During the summer months
 - 2) During the winter months
 - 3) When they are directly between the earth and the sun
 - 4) For a short time after sunset or before sunrise
34. Fig 3 For an observer at X, what is the phase of the moon when it is in the position represented at D?
- 1) New moon
 - 2) Full moon
 - 3) First quarter
 - 4) Second quarter

Sample F.¹⁰

Water takes up oxygen from the air in varying amounts. Cold water will take up small quantities of oxygen while warm water takes up almost

⁸ Truman L Kelley, Giles M Ruch, and Lewis M Terman, *Stanford Achievement Tests, Elementary Science*, Advanced Battery Published by World Book Co., 1940

⁹ Alvin W Schindler, *The 1940 Iowa Every-Pupil Test in General Science* Published by University of Iowa, 1940

¹⁰ John G Zimmerman and Richard E Watson, *Cooperative Science Test for Grades 7, 8, and 9*, Form R Published by Cooperative Test Service, 1941.

none. Running water will dissolve (that is, take up) more oxygen than standing water. Water in which plants are growing contains much oxygen because the green plants give off oxygen in the process of photosynthesis. When there is not enough light for plants to manufacture food, they do not give off oxygen but consume it in respiration. Water animals also use oxygen in respiration so that the amount of oxygen found in water is always changing. The oxygen content of an aquarium changes from day to day and from hour to hour and is different even at different levels in the aquarium.

13. Standing water takes up

13-1 more oxygen than running water.

13-2 as much oxygen as running water.

13-3 less oxygen than running water.

13-4 a great deal of oxygen.

13-5 no oxygen.

..... 13()

14. The manufacture of food by plants requires water and carbon dioxide. It also requires

14-1 light

14-2 a small amount of oxygen.

14-3 a large amount of oxygen.

14-4 a high temperature.

14-5 a very low temperature.

..... 14()

Matching Exercises. Three samples of the matching test are given below. The first illustrates the common form of item based on word-phrase relationships, the second, illustrating an identification test, requires the matching of parts of the digestive tract and their pictorial representation, and the third is a matching unit having some elements in common with the multiple-choice form.

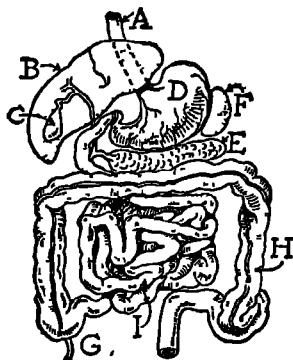
Sample G (Incomplete—20 words and questions in test).¹¹

A. WORDS

B. QUESTIONS

amplifier	1. Used to view light from an incandescent body_____
barometer	2. Attracts iron or steel _____
caisson	3. Used to look at heavenly bodies_____
camera	4. Used to measure heat intensity _____
compass	5. Used in measuring time _____
hydrometer	6. Used to keep liquids at a constant temperature_____
lever	7. Used to magnify the electric impulse_____
flax	8. Used to decompose sewage _____

¹¹ Robert K. Speer, Lester D. Crow, and Samuel Smith, *National Achievement Tests, General Science, Grades 7-9*. Published by Acorn Publishing Co., 1939

Sample H.¹²

In this diagram of the digestive tract:

- | | | |
|----------|---------------------------------|----------|
| <i>a</i> | The small intestine is lettered | <i>a</i> |
| <i>b</i> | The œsophagus is lettered | <i>b</i> |
| <i>c</i> | The liver is lettered | <i>c</i> |
| <i>d</i> | The stomach is lettered | <i>d</i> |
| <i>e</i> | The pancreas is lettered | <i>e</i> |

Sample I.¹

- | | | |
|---------------|-----------------------------------|------|
| 1 Heart | 1. Circulatory system | 1() |
| 2 Kidney | | |
| 3 Lung | 2. Excretory system | 2() |
| 4 Stomach | | |
| 5 Spinal cord | 3. Nervous system | 3() |
| 1 Mendel | 7. Laws of biological inheritance | 7() |
| 2 Pasteur | | |
| 3 Hooke | 8. Germ theory of disease | 8() |
| 4 Burbank | | |
| 5 Carrel | 9. Plant breeding | 9() |

IV. INFORMAL OBJECTIVE TESTING IN ELEMENTARY SCIENCE

Objective items of the types illustrated on pages 429 to 432 have been used quite widely by informal objective test workers in the evaluation of the more intangible outcomes of science instruction. Furthermore, they have also used rather complex adaptations of the common item forms. There seems excellent reason to believe that much of the most significant recent testing in the science field has been done by informal objective testing methods. Perhaps three major reasons why these techniques have not yet entered

¹² Ruch and Popenoe, *op. cit.*

¹³ O E Underhill and S R Powers, *Cooperative General Science Test*, Form Q. Published by Cooperative Test Service, 1940.

into the standardized testing field below the college level are that (1) they have been applied in the main to situations requiring a variety of applications of science knowledges and skills, (2) they are rather difficult to construct and standardize, and (3) they frequently run to considerable length.

Space does not permit many illustrations of this type of approach to the measurement of scientific knowledges and abilities. Although the illustrations given are mainly from testing for the junior or the senior high school levels, it does not follow that these techniques are applicable only to instructional outcomes of greater complexity than those of the elementary school. It means, rather, that the most significant work of this type has so far been done at the high school and college levels. The same techniques can be, and in some cases have been, adapted to the elementary sciences.

Measurement of Broad Instructional Outcomes. An illustration of the application of evaluative techniques to the measurement of broad instructional outcomes of elementary science is given by LuPone¹⁴ for a unit on machines and their implications. The following chart shows the relationships LuPone established between pupil outcomes, tools used in evaluating the outcomes, and illustrations of how the tools might be applied. Several illustrations of item types which cannot be reproduced here are also given in the reference. Observation of the tools used in the paralleling areas of evaluation and of the typical illustrations given in the chart should indicate to the student the possibilities of measurement employing a variety of techniques.

EVALUATION CHART

<i>Areas of Evaluation</i>	<i>Tools Used</i>	<i>Typical Illustrations</i>
<i>Concepts</i>		
1 Our ways of living are affected by the use of machines	Classroom tests.	Problems.
2. Man's intelligent endeavor has been a factor in our present civilization.	Work sheets.	

¹⁴ O J LuPone, "Evaluating the Intingibles in Elementary Science." *School Science and Mathematics*, 39 754-59, November 1939

EVALUATION CHART

<i>Areas of Evaluation</i>	<i>Tools Used</i>	<i>Typical Illustrations</i>
3. Our society is affected by inventions. 4. Machine power is more efficient than man power. 5. The era of machines has implications which are social. 6. Human energy can be conserved by use of machines.	Parent interviews. Anecdotal records of pupil behavior based upon teacher observation. Pupil logs or diary of the child's experience during the area of work.	
<i>Knowledges</i>		
1. A knowledge that machines can do work more quickly, more easily, better, than man power. 2. Machines are a combination of two or more simple machines. 3. "Machines give us more leisure time for recreation.	Performance tests. Anecdotal records. Pupil logs and pupil diaries.	Experiments with simple machines. Problems.
<i>Attitudes and Appreciations</i>		
1. Pupils have respect for people who develop more efficient ways of doing. 2. An appreciation that the standard of living is higher because of machines 3. A recognition that society changes through the effect of inventions.	Comparison of pre-tests and final tests covering attitudes and appreciation. Classroom tests. Observations.	Statements about the unit of work by which the child can express what he believes.

EVALUATION CHART

<i>Areas of Evaluation</i>	<i>Tools Used</i>	<i>Typical Illustrations</i>
<i>Overt Behavior</i>		
1. A desire to visit machines at work.	Excursions.	A visit to a nearby project under construction.
2. A desire to read and send for material about machines	Anecdotal records.	
3. A desire to use simple machines.	Writing for source materials.	
<i>Skills</i>		
1. Use and construct simple machines.	Performance tests.	The construction of simple machines.
2. Organization of materials.	A classroom test based upon skills devised by the teacher.	
3. Manipulation of apparatus.		

An informal, semi-objective test for teaching more than for testing purposes was devised by Davis¹⁵ for use in eighth- or ninth-grade science courses in the measurement of other than largely factual instructional outcomes. The following reproduction of the instructions to pupils and of the first paragraph of the selection to be read and evaluated by the pupils will serve to show the nature of the instrument.

TO THE PUPIL

Here is a test which I think you will find quite different from any you have ever taken. It is a story about Johnny Jones. He was quite an active boy, but sometimes he was a poor scientist. Some of his friends and the members of his family may not have been good scientists either. Whenever you find something in the story which does not agree with what you think good science means, put a pair of parentheses () around the sentence or part of a sentence where you find this. Next, at the border of the paper beside the error, write in the correct letter from the following list.

S means that Johnny or some one else was superstitious.

¹⁵ Warren M. Davis, "A Science Test Designed to Teach and Measure Outcomes Other than Memorization of Factual Information" *Science Education*, 23:371-72, December 1939.

- D means that what was being done or had been done was dangerous.
 O means that statements are being taken or have been taken for truth without any proof being offered
 J means something unscientific for reasons other than S, D or O If you use the letter J be prepared to tell the class what was wrong with the story at the point where you use this letter.

Now go on with the story

JOHNNY'S DAY

Johnny Jones woke from a sound sleep one morning and noticed that the sun was already shining in his window Without looking where he was going he jumped to the floor and started gathering up his articles of clothing to put them on. Suddenly he stopped and said, "Shucks, it's Saturday, no need for me to hurry But it might just as well be a school day," he went on as he looked out of the window, "it's sure to rain today. Old man Smith said this was a wet moon"

The remaining parts of the selection, running to perhaps 1100 words, included many additional evidences of behavior or reasoning illustrative of the types of situations covered by the S, D, O, and J methods of marking the selection. One point of credit was assigned for each pair of parentheses placed approximately in the correct position and an additional point of credit was assigned for each pair of parentheses accompanied by the proper identifying letter.

Measurement of Scientific Attitude. Noll lists the following six abilities as essential to the scientific attitude: (1) accuracy in all operations—calculation, observation, and report, (2) intellectual honesty, (3) open-mindedness, (4) the habit of looking for natural causes, (5) the habit of suspended judgment, and (6) the habit of criticism.¹⁶ Although he admits that other habits might be included in such a list, he states that a person who met all of the conditions listed above would possess the scientific attitude and would also be highly unique.

Suggestions concerning how each of these six essentials of scientific attitude can be measured informally are also presented by Noll.¹⁷ Some of his illustrations are reproduced to show techniques useful in measuring scientific attitude.

¹⁶ Noll, *op cit.* pp. 25-26.

¹⁷ *Ibid* pp 34-37

- (1) Accuracy in calculation—arithmetic examples
Accuracy of observation and report—questioning a pupil concerning the characteristics of an animal picture, plant, or diagram.
- (2) Intellectual honesty.
T F When a pupil makes a poor mark in an examination it is usually because he is not well or he was up late the night before
T F It is perfectly justifiable not to pay one's fare on a bus or street car if the conductor doesn't come around to collect it.
- (3) Open-mindedness.
T F All Indians are dirty.
T F College professors as a rule would be failures in any line of work but teaching.
- (4) Cause and effect relationships
T F Finding a horseshoe means that one will have good luck.
T F Giraffes have such long necks because through many generations they have been stretched a little longer each time.
- (5) Suspended judgment
T F My neighbor is away from home most of the time. He must be a traveling salesman
T F Mr. Jones bought a new car last week. He must have had an increase in salary
- (6) Criticism
T F One can always accept as true what is printed in a book.
T F If my science teacher says a thing is so, it must be so.*

Another approach to the measurement of similar types of outcomes is presented by Davis, who gives the directions to pupils and a few sample items from a test for measuring knowledge of cause and effect relationships.¹⁸ Students were asked to indicate by the use of the appropriate letter of the following

- A—If the first occurrence is practically the sole cause of the second.
- B—If the first occurrence is one of a number of the important contributing causes of the second
- C—If the first occurrence contributes only slightly to the second.
- D—If both occurrences are results of the same general cause or causes.
- E—If the first occurrence bears no causal relationship to the second.

¹⁸ Ira C. Davis, "The Measurement of Scientific Attitudes" *Science Education*, 19 117-22, October 1935

their reactions to such items as these

1. The sun shines on the earth, the earth is warm
2. A boy often picked up toads, the boy had warts on his hands.
3. The light of lightning, the accompanying thunder.
4. The ignition switch of an auto is turned on, the motor starts running
5. A rising column of air was cooled, a cloud formed.

Davis also gives similar illustrations from a test designed to measure ability to distinguish between fact and theory.¹⁹ The appropriate letter from this list

- A—Some are statements of well established facts which are always true
- B—Others may be statements of well established theories which are generally accepted.
- C—Others may be statements of theories which are questioned by some (many) authorities
- D—Others may be statements of popular beliefs which are not supported by evidence

was to be used in responding to each of these statements

1. A disease is a punishment for some particular moral wrong.
2. Air is composed of molecules
3. The pressure in water varies with the depth.
4. Heating the molecules in air increases their speed.
5. A high forehead indicates high intelligence

Measurement of Superstitious Beliefs. Zapf presents a technique for measuring the manner in which pupils actually behave in situations to which well-known superstitious beliefs apply.²⁰ Pupils were placed in a closed room, where they opened boxes in which were found directions for their subsequent action asking that they go contrary to widely held superstitious beliefs. The extent to which they performed the actions was taken as an indication of the degree to which they were not governed in their behavior by these beliefs. Such situations as breaking a mirror, walking under a ladder, and opening an umbrella indoors, all relatively simple performances, were among the twelve used in the test. Although all thirty-two pupils tested in these situations had previously indicated that they did not believe in the super-

¹⁹ Ibid

²⁰ Rosalind M. Zapf, "Superstitious Beliefs" *School Science and Mathematics*, 39 54-62, January 1939

stitutions, only two pupils went contrary to all twelve superstitions and two pupils acted superstitiously in five of the twelve situations.

Controlled Observation. A controlled observation procedure for use in elementary school science was worked out by West.²¹ He devised a tabulation sheet and observation procedures of too great complexity for presentation here for use in the classroom evaluation of the dynamic and the performance factors of pupil behavior. Inquiry, critical-mindedness, open-mindedness, generalizing, recognition of achievements of thinking, scientific problem-attack, recognition of interpretations of natural phenomena, and cause and effect relationships were listed as dynamic factors, while responsibility, voluntary activity, initiative, application of experience, self-appraisal, resourcefulness, skills, special abilities, work habits, and miscellaneous were listed as performance factors. His recommendation is that this objective type of observational procedure be used to supplement but not to supplant the measurement techniques in common use in the classroom.

V. DIAGNOSIS AND REMEDIAL TEACHING IN ELEMENTARY SCIENCE

Limitations of Diagnostic and Remedial Techniques in Science. Diagnostic procedures and remedial work in the field of science instruction are not highly developed. While certain of the available tests may show pupils to be deficient in some specific phase of science information, the majority of such tests do not point out the causes of the deficiencies. Practically all that can be done by way of diagnosis is in connection with certain skills which appear to be basic to the study of science.

The study of science involves the comprehension of a language peculiar to the subject. Reading of scientific content is apt to be difficult. Thus, poor reading ability may form the basis of poor accomplishment in the subject.

²¹ Joe Young West, *A Technique for Appraising Certain Observable Behavior in Science in Elementary Schools* Contributions to Education, No. 728 Teachers College, Columbia University, New York, 1937

Diagnosis of reading abilities of the work-study type, accompanied by remedial instruction designed to overcome the weaknesses revealed, is one of the prerequisites to satisfactory progress in the study of the sciences. Laboratory work may call for many new abilities and techniques.

Future of Diagnosis in Science. There is considerable promise for the future of diagnosis and remediation in the sciences through further development of the evaluation techniques illustrated in the preceding section of this chapter. The attempt so far has been more upon the construction of valid evaluation procedures for the less tangible outcomes of instruction than upon diagnostic values of the techniques. The writers believe, however, that constructive diagnostic and remedial procedures may well grow out of this new approach to the measurement of ability in the sciences.

TOPICS FOR DISCUSSION

1. Why is objective measurement in the sciences not highly developed?
2. Enumerate and evaluate the chief general aims of education which presumably are met by science instruction.
3. In your opinion is the need for a unified course in science in the intermediate grades any less serious than it is in the social studies?
4. What are the four most important measurable outcomes of instruction in science?
5. Examine the possibilities of measuring the acquisitive skills in science and specify a number of techniques in each. Would such a list parallel the types found in the social studies?
6. Make a list of specific outcomes of instruction in general science.
7. What appears to be the present tendency with respect to nature study and physiology and hygiene in the elementary science field?
8. Suggest some of the objective item types useful in science testing and illustrate them with items of your construction.
9. Discuss and evaluate the informal objective test approaches to the measurement of some of the more intangible outcomes of science instruction.

SELECTED REFERENCES

- Arnold, Dwight L., "Testing Ability to Use Data in the Fifth and Sixth Grades." *Educational Research Bulletin*, 17 255 ff, December 7, 1938
- Bedell, Ralph C., "A Method of Diagnosis and Remedial Treatment in General Science." *Science Education*, 13 260-66, May 1929.
- Buckingham, Guy E., and Lee, Richard E., "A Technique for Testing

- Unified Concepts in Science." *Journal of Educational Research*, 30 20-27, September 1936.
- Croxton, W. C, *Science in the Elementary School*. New York McGraw-Hill Book Co, Inc, 1937
- Curtis, Francis D, "Diagnosis and Remedial Treatment in the Field of Science." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XVI, pp. 331-45 Bloomington, Ill Public School Publishing Co, 1935.
- Curtis, Francis D, *A Digest of Investigations in the Teaching of Science in the Elementary and Secondary Schools*. Philadelphia P. Blakiston's Son and Co, 1926.
- Curtis, Francis D, *Second Digest of Investigations in the Teaching of Science*. Philadelphia P. Blakiston's Son and Co, 1931.
- Curtis, Francis D, *Third Digest of Investigations in the Teaching of Science*. Philadelphia P. Blakiston's Son and Co., 1939.
- Davis, Ira C., "The Measurement of Scientific Attitudes." *Science Education*, 19 117-22, October 1935.
- Davis, Warren M, "A Science Test Designed to Teach and Measure Outcomes Other than Memorization of Factual Information." *Science Education*, 23 371-72, December 1939.
- Downing, Elliot R, "Some Results of a Test on Scientific Thinking." *Science Education*, 20 121-28, October 1936.
- Frutchey, Fred P, "Evaluation in Elementary School Science" *Educational Method*, 16 422-26, May 1937.
- Frutchey, Fred P, "Testing for Application of Scientific Method" *Educational Method*, 15 427-32, May 1936.
- Hart, E. H, "Measuring Critical Thinking in a Science Course" *California Journal of Secondary Education*, 14 334-38, October 1939.
- Hawkes, Herbert E, Lindquist, F. F, and Mann, C R (Editors), *The Construction and Use of Achievement Tests*, Chapter V. Boston: Houghton Mifflin Co., 1936
- Irwin, Manley E, "The Measurement of Nature Study in the Primary Grades in the Detroit Public Schools" *Science Education*, 15 23-32, November 1930.
- Krug, Edward A, "A Cooperative Approach to Evaluation." *California Journal of Secondary Education*, 14 346-52, October 1939
- LuPone, O. J, "Evaluating the Intangibles in Elementary Science." *School Science and Mathematics*, 39 754-59, November 1939.
- Maller, J. B., "Superstition and Education" *Encyclopedia of Educational Research*, pp 1186-90 New York The Macmillan Co, 1941.
- Noll, Victor H, *The Teaching of Science in Elementary and Secondary Schools*. New York Longmans, Green and Co, 1939.
- Powers, Samuel Ralph, "Contributions of Research to Special Methods: Natural Science" *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XI, pp. 135-46. Bloomington, Ill.: Public School Publishing Co., 1938.

- Science in General Education*. Report of the Committee on the Function of Science in General Education, Commission on Secondary School Curriculum, Progressive Education Association, Chapter IX. New York D. Appleton-Century Co., Inc. 1938.
- Smith, Henry Lester, and Wright, Wendell William, *Tests and Measurements*, Chapter XIII. New York Silver, Burdett and Co., 1928
- Stewart, A W, "Measuring Ability to Apply Principles." *School Science and Mathematics*, 35 695-99, October 1935.
- Tiegs, Ernest W, *The Management of Learning in the Elementary Schools*, Chapter XIII. New York Longmans, Green and Co., 1937.
- Tyler, Ralph W, *Constructing Achievement Tests*, pp. 24-52. Columbus, Ohio Ohio State University, 1934.
- Webb, L W, and Shotwell, Anna Markt, *Testing in the Elementary School*, Chapter XVI New York Farrar and Rinehart, Inc., 1939.
- West, Joe Young, *A Technique for Appraising Certain Observable Behavior in Science in Elementary Schools*. Contributions to Education, No 728. New York Teachers College, Columbia University, 1937.
- Wrightstone, J. Wayne, *Appraisal of Newer Elementary School Practices*, Chapter XI New York Bureau of Publications, Teachers College, Columbia University, 1938.
- Zapf, Rosalind M., "Superstitious Beliefs." *School Science and Mathematics*, 39 54-62, January 1939.

CHAPTER XIX

MEASUREMENT IN THE FINE ARTS

This chapter presents a discussion of the following possibilities of measurement in the fine arts.

- a.* Social and educational significance of the arts.
- b.* Educational emphasis on art and music.
- c.* Basic elements of musical talent
- d.* Measurement of musical accomplishment.
- e.* Measurement of art appreciation.
- f.* Measurement of artistic ability.

Introduction. The objective measurement of achievement in the fine arts is a relatively recent accomplishment. In fact, it is so recent that there is still an echo of protest from certain conservative artists that artistic production does not lend itself to measurement. In spite of this feeling, however, recent years have seen much progress in these fields. This is as it should be, for certainly in these cultural subjects is to be found much of the best that our educational program affords. With the trend of recent years in the direction of greater individual leisure for the cultural pursuits, the need for a better understanding of the content, aims, and methodology of these artistic subjects is greater than ever before.

There is perhaps a certain advantage in the fact that developments in the measurement of the fine arts have taken place slowly and rather recently. In general, research techniques have improved, with the net result that the problems of measurement in these fields have been more critically analyzed and attacked with more refined instruments. The very critical and carefully controlled research work of such men as Seashore, Schoen, Kwalwasser, and Dykema in the psychology and pedagogy of music, and the work of Thorndike, Ayer, Meier, Manual and Whitford in the field of art are evidences of the influence of this point of view.

I. MEASURABLE QUALITIES IN MUSIC

Music Talent and Achievement. Measurement in music takes two major lines of approach. The first is the determination of basic aptitudes. Here, as in other subjects, the techniques and instruments used are psychological. Such instruments have been mentioned previously in this volume as tests of specialized intelligence, since they have to do with the determination of inherited tendencies to respond in certain ways to specific types of musical stimuli. Accomplishment in music depends to such a large degree upon the existence of aptitude that this phase of measurement must be emphasized. The mere existence of aptitude in music is in no sense an index to musical accomplishment, however. The second approach to the problem is pedagogical and is based upon the use of achievement tests for the threefold purpose of measuring the knowledges, skills, and appreciative aspects acquired as a result of training. As Kwalwasser¹ points out: "Regardless of the talent possessed, one must have the will to succeed or little is attained. . . There are a vast number of reasons why an individual of superior endowment may realize but a very small return on his native musicianship."

Aims and Outcomes of Music Education. The 1921 report of the *Educational Council of the Music Supervisors' National Conference*² presented a standard course of study in music, setting forth the aims and outcomes of music education in a form which has not been significantly improved upon since that time. While the aims and attainments are listed for each grade in the statement, the general aims appropriate for the end of the sixth grade only are summarized here. All aims and outcomes of prior grades are incorporated in this summary:

SIXTH-GRADE AIMS

- a To continue the development of free and beautiful singing of songs.
- b To acquire an increasingly wide musical experience
- c To develop increasing power of eye and ear in correlation

¹ Jacob Kwalwasser, *Tests and Measurements in Music* C C Birchard Co., Boston, 1927

² Report of Educational Council of the Music Supervisors' National Conference. National Education Association, Washington, D C, 1921

- d.* To develop power to listen for musical beauty as well as for musical knowledge.
- e.* To develop increased power to sing at sight.
- f.* To establish two-part singing.
- g.* To develop increasing practical knowledge of the tones of the chromatic scale and power to use them.
- h.* Extension of knowledge of the tonal and rhythmic material of music appropriate to the sixth year.
- i.* To develop a fair degree of power to sing unison songs at sight with words, and an elementary degree of power to sing two-part songs at sight with words
- j.* To begin the development of three-part, treble-voice singing.
- k.* To develop ability to deal practically with the minor mode.

SIXTH-GRADE ATTAINMENTS

- a.* Ability to sing well, with enjoyment, at least 30 unison, two-part, and three-part songs, some of which shall be memorized.
- b.* Ability of 90 percent of the pupils to sing individually, freely, correctly, and without harmful vocal habits not less than ten of the songs sung by the class as a whole.
- c.* Ability to sing at sight, using words, a unison song of hymn-tune grade; or using syllables, a two-part song of hymn-tune grade, and the easiest three-part songs, these to be in any key; to include any of the measures and rhythms in ordinary use, to contain any accidental signs and tones easily introduced, and in general to be of the grade of folk songs such as "The Minstrel Boy." Also knowledge of the major and minor keys and their signatures.
- d.* Ability of at least 30 percent of the pupils to sing individually at sight music sung by the class as a whole.
- e.* Ability to appreciate the charm of design in songs sung, to give an account of the salient features of structure in a standard composition, after a few hearings of it, to identify at least the three-part song form from hearing, to recognize and give titles and composers of not less than 20 standard compositions studied during the year.

II. MEASUREMENT OF MUSICAL TALENT

Measurement of Basic Musical Talent. Tests of musical aptitude are designed to measure those innate musical capacities which constitute the individual's musical inheritance. Aside from the sheer physical endowment which certain types of musical expression demand, there are certain more or less psychological factors which determine an individual's musical talent. The identification of these factors calls for an unusually critical analysis. Without doubt one

of the most extensive research programs ever undertaken for the purpose of isolating the elements of native capacity in a special field is that undertaken by Seashore and his students in connection with the development of his *Tests of Musical Talent*.

These tests, six in number, are designed for use not earlier than the fifth grade. Experience indicates that this is practically the earliest age at which such group measures may be taken, and that it is a suitable age at which to begin making serious arrangements for a musical education if it appears desirable. The six elemental abilities measured by the tests in their present form are (1) pitch, (2) loudness, (3) time, (4) timbre, (5) rhythm, (6) tonal memory. The test stimulus to which the pupil responds is supplied by means of three twelve-inch double-faced records for phonographic reproduction. The tests may be administered to groups, the size of the group depending somewhat on the acoustical qualities of the room. Naturally the stimulus must be heard clearly at all times.

According to the author's own statement,

These measures present the following characteristics: they are based on a scientific analysis of musical appreciation and performance, they deal with elements which function in all music; they are standardized for content so that alternate or new series are not needed, they give quantitative results which may be verified to a high degree of certainty, they are economical in that expensive instruments are replaced by phonograph records, they may be used with any language and at any racial or cultural level, they are simple and as nearly self-operative as possible, they are designed for group measurements, they are interpreted in terms of established norms³

The *Kwalwasser-Dykema Music Tests* are quite similar in form and function to the *Seashore Musical Talent Tests*. The ten tests, designed for use in grades four to twelve, require five phonograph records. The elements measured by the alternate-response technique are: (1) tonal memory, (2) quality discrimination, (3) intensity discrimination, (4) tonal memory, (5) tone discrimination, (6) rhythm dis-

³ Carl E. Seashore, Don Lewis, and Joseph Saetveit, *Manual of Instructions and Interpretations for the Seashore Measures of Musical Talent*. RCA Manufacturing Company, Inc., Educational Department, Camden, New Jersey, 1939

crimination, (7) pitch discrimination, (8) melodic taste, (9) pitch imagery, and (10) rhythm imagery.

Measurement of Musical Memory. Quite in contrast with the two tests of musical talent discussed above is the *Drake Musical Memory Test*, which measures musical aptitude by an entirely different technique. The test is designed for persons of any age above seven whether or not they have had musical training. The subject listens to twelve melodies played in their proper form or with variations in key,

EXCERPTS FROM SCORE SHEET, DRAKE MUSICAL MEMORY TEST⁴

1. There are 12 trials of entirely different melodies
2. Listen carefully to the first melody in each trial and remember it.
3. Listen to what is played next and compare it to the first melody to determine
 - a if it is exactly the same as the first melody, - if so record S
 - b if it is the same melody played in a different key, - if so record K.
 - c if the time has been changed, - if so record T
 - d if any notes have been changed, - if so record N

S=exactly the SAME melody
K=change of KEY

T=change of TIME
N=change of one or more NOTES

Practice exercise No 1

Practice exercise No 2

4. Record your answers in the score form given below
5. Each trial will be announced by number. When you hear a number announced you will know that a new melody is to be played to which all melodies that follow, in that trial, are to be compared
6. Record your answer during the short pause between each melody. Just time enough will be given to write your answer
7. There is never more than one kind of change in any one comparison.
8. Fill in every square. Make the best judgment you can for each comparison.
9. Write clearly with capital letters
10. In each trial, listen to the first melody. Wait until more is played and record whether it is the same, or if a change has been made in time, key, or notes

IF THERE IS ANYTHING YOU DO NOT UNDERSTAND ASK ABOUT IT NOW.

Remember—
S=SAME
K=KEY change
T=TIME change
N=NOTE change

1												
2												
3												
4												
5												
6												
7												
8												
9												
10												
11												
12												

Errors

SCORED BY

TOT -
ERRORS=FINAL SCORE

⁴ Raleigh M Drake, *Musical Memory Test Score Sheet* Published by Public School Publishing Co, 1934

time, or notes. He records his responses to each of the 54 trials on a special record sheet to show whether he recognizes the nature of the difference, if any, between the melody itself, which is played first, and the various versions of it which follow. The reproduction of the directions and of the response section of the score sheet on page 447 shows the manner in which the test is given and the manner of recording responses.

III. MEASUREMENT AND REMEDIATION IN MUSICAL ACHIEVEMENT

The knowledge, skill and appreciative outcomes of music instruction are measured by a variety of tests of the pencil-and-paper variety. The majority of these instruments appear to measure the knowledge and skill objectives quite adequately, but appreciative outcomes are largely neglected. This is not surprising, because of the fact that appreciations are almost impossible to define and extremely difficult to measure.

The *Beach Standardized Music Tests* were among the real pioneers in the measurement of musical achievement. Many of the elements measured by these tests are recognized among the tests of more recent development. The following qualities are scheduled for measurement by the test:

1. Knowledge of essential facts of musical notation
2. Ability to hear and distinguish the component parts of music, namely the elements of time and tune both in isolated form and in melodies.
3. Aural recognition of the structural elements of music fundamentally necessary for intelligent appreciation.
4. Pitch discrimination
5. Musical memory
6. Sight-singing through indirect methods.
7. The writing of music

Measurement of Musical Knowledge. Tests of musical knowledge are variously concerned with musical symbols and terms, time and key signatures, note and rest values, syllables, instrumentation of the orchestra, musical form, and the history and biography of music. Samples are given below to illustrate the measurement techniques rather commonly used. Multiple-choice and simple recall items and

matching exercises appear to be most common among the testing techniques used, although the true-false item is used occasionally. The following samples are somewhat representative of the content of various tests.

Simple Recall Items. Two illustrations of this item type are given below. The first represents the common type of simple recall and the second illustrates a specific adaptation to testing of musical knowledge.

Sample A.⁵

COMPOSERS OF FAMOUS COMPOSITIONS



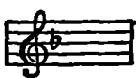
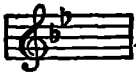
Directions: Below are the names of famous compositions. On the lines at the right you are to write the *name of the composer* of each. The sample is marked as it should be.

Sample: The Elijah *Mendelssohn*

- | | |
|----------------------------|-----|
| 1. March Slav | |
| 2. To a Wild Rose | . |
| 3. The Unfinished Symphony | . . |
| 4. Liebestraum | . |

Sample B.⁶

In the major key signatures below determine what the name of each one is. Find that name above, take its number, and write it in the blank at right of each one, as shown at a. Ready! Go!

a.		No. <u> </u>	a.
b.		No. <u> </u>	b.
g.		No. <u> </u>	g.
h.		No. <u> </u>	h.

True-False Items. The true-false item is illustrated below as applied to general knowledge concerning instrumentation. It can be used in many other ways in measuring musical knowledge.

⁵ Jacob Kwalwasser, *Kwalwasser Test of Music Information and Appreciation*. Published by Bureau of Educational Research and Service, University of Iowa, 1927.


⁶ Clara J. McCauley, *McCauley Examination in Public School Music*. Published by Jos. E. Avent, 1933.



Sample C.⁷

1. () The viola is an alto horn.
2. () Violins are frequently employed in brass bands.
3. () The first violin section is seated to the left of the conductor
4. () The harpsichord is one of the predecessors of the piano.

Multiple-Choice Items. The two following illustrations show adaptations of the multiple-choice item to the measurement of knowledges concerning time signatures and note values. The same form is readily used for rest values as well. Directions are not reproduced here for the sample items.

Sample D.⁸

1.  The time signature is $\frac{2}{4}$ $\frac{3}{4}$ $\frac{4}{4}$ $\frac{2}{8}$ $\frac{3}{8}$ 1

1.  The note needed is  1

Matching Exercises. An example of a matching exercise in which one set of items is used several times each is given below. Balanced matching sets are also used in various music tests.

Sample E.⁹

TEST 1. The Way Musical Instruments Are Played

Directions: Below is a list of musical instruments. They are not all played in the same way. In the blank space opposite each instrument write the letter

A—for each one played by blowing

C—for each one played with a bow

B—for each one played by plucking, or picking

D—for each one played by striking, or shaking

The sample is marked as it should be

Sample: Ukulele B (because it is played by plucking.)

Begin here.

- | | | | |
|----------------|------------------|-------------------|----------------------|
| 1. Cornet_____ | 7. Trumpet_____ | 13. Trombone_____ | 19. E-flat Alto_____ |
| 2. Banjo_____ | 8. Mandolin_____ | 14. Guitar_____ | 20. Harp_____ |

⁷ Kwalwasser, op cit

⁸ Jacob Kwalwasser and G M Ruch, *Kwalwasser-Ruch Test of Musical Accomplishment* Published by Bureau of Educational Research and Service, University of Iowa, 1924

⁹ Glenn Gildersleeve and Wayne Soper, *Musical Achievement Test* Published by Bureau of Publications, Teachers College, Columbia University, 1933

Measurement of Musical Skills. Among the musical skills most commonly measured by various tests are detection of pitch and time errors and recognition of melodies. Illustrations of each are given below. The first represents a type of matching situation and the second a recognition form of item measuring ability to detect errors.

Sample F.¹⁰

- (e) Below are printed the opening strains of five familiar melodies. After reading or humming them one by one, select the title of each from the list of answers below. Then place the corresponding number in the square at the right of each melody. The sample is correct.

List of Answers

Silent Night	5	America the Beautiful
Old Black Joe	6	Auld Lang Syne
Santa Lucia	7	Star Spangled Banner
Home Sweet Home	8	America

Sample :

The image shows three staves of musical notation for the song 'The Rose Tree'. The first staff is in treble clef, 4/4 time, with a key signature of one flat (B-flat). It contains the melody for the first line of the song. The second staff is in treble clef, 4/4 time, with a key signature of one flat. It contains the melody for the second line of the song. The third staff is in treble clef, 4/4 time, with a key signature of one flat. It contains the melody for the third line of the song. To the right of each staff is a square box for a response, with the number '5' in the first box.

Measurement of Music Appreciation. Only one test of music appreciation, the *Kwalwasser Test of Music Information and Appreciation*, is known to the authors. Its approach is mainly through the testing of knowledges, many of which unquestionably carry appreciative values with them. However, it does not appear that appreciations are measured directly, if, indeed, they can be measured in that manner. In view of the modern emphasis upon music appreciation for all pupils, it is unfortunate that the appreciative types of outcomes are not subject to satisfactory measurement.

¹⁰ Frank A. Beach, *Beach Music Test* (Revised) Published by Bureau of Educational Measurements, Kansas State Teachers College, Emporia, 1938

Sample G.¹¹

TEST 3. DETECTION OF PITCH ERRORS IN A FAMILIAR MELODY

DIRECTIONS. The song "America" is written below. One measure has been crossed out because the melody is wrong. Five other measures are wrong. Hum over the melody to yourself and cross out all five wrong measures.

Begin here



Remedial Materials in Music. Remedial instruction in music consists mainly in the application of further practice on the specific skills desired. Musical aptitudes, of course, do not lend themselves to corrective treatment. Certain of the phases of musical accomplishment respond to additional, or different, treatment. Thus remedial teaching in this field becomes mainly a process of identifying certain habits and informations and applying to them the required further drill and instruction. As a matter of fact, there is available almost no material of definitely remedial nature in the field of music.

IV. CHARACTERISTICS AND AIMS OF ART EDUCATION

General Trends in Art Education. The theory of art for art's sake which dominated the field of art education for many years has largely given way of late to the theory that all pupils should receive an opportunity in art courses to develop a sensitivity to beauty and critical taste in evaluating art objects. Hence, art is no longer thought of as a field only for the talented few. Creative self-expression, especially in the lower grades, and correlation of art with other activities of the school are important modern trends. These trends involve the use of a wide variety of art materials in

¹¹ Kwalwasser and Ruch, op cit.

the classroom. Extension of the content of art education courses beyond the drawing and painting which largely constituted the curriculum in the past particularly to industrial arts is another trend worthy of note. Last, and perhaps most important, the appreciative aims of art education have increasingly come to the front.¹²

General Outcomes of Art Education. Three general outcomes of art education appear to be of major importance: (1) information, (2) appreciation, and (3) expression.¹³ It is quite probable that art appreciation is not necessarily taught, although real appreciation may be considered to rest to a large degree upon the broader aspects of information. There will still remain something in the truly artistic product which sheer information does not entirely explain. The third major objective might be better expressed as exploration. Not many potentially great artists are discovered in the elementary school classroom, but practically all the great artists there are have come up through this avenue. Not everyone can express himself effectively in artistic form, but everyone has a right to explore for himself the fields of human expression in the hope that his own hidden talent may be uncovered. Art talent and achievement tests have distinct contributions to make in this field.

Specific Outcomes of Art Education. The elementary school art course faces the responsibility for bringing to the child a four-fold artistic experience. These categories of experience were suggested by Kirby¹⁴ in a discussion of aims and tendencies in art education. The first is the *graphic* experience which expresses itself in representative drawing, illustrative and imaginative drawing, nature drawing, and other related forms. The second is the *thoughtful* experience involving the constructive, decorative phases of artistic expression. The third experience involves the *acquisition*

¹² Robert S. Hilpert, "Changing Emphases in School Art Programs" *Art in American Life and Education* Fortieth Yearbook of the National Society for the Study of Education, pp. 452-53. Public School Publishing Co., Bloomington, Ill., 1941.

¹³ Walter H. Klar, Leon L. Winslow, and C. Valentine Kirby, *Art Education in Principle and Practice* Milton Bradley Co., Springfield, Mass., 1933.

¹⁴ C. Valentine Kirby, "Aims and Tendencies" *Pennsylvania School Journal*, 77 501-2, April 1929.

of *motor skill* in expression. The fourth is the *emotional* experience which involves the appreciation of the arts.

A somewhat more specific expression of outcomes of instruction in art is given in the accompanying outline of specific outcomes adapted from a course of study covering the first six years of the elementary school. It will be noted that

OUTCOMES OF ART INSTRUCTION ¹⁵

- A. Fruitful knowledge
 - Functional information
 - Practical relation of art to everyday life (clothing, home, town or city, etc)
 - Understanding of elements and principles of art and their adaptation to everyday use
 - Knowledge of construction and industrial processes involved in art training
 - Acquaintance with art of other countries
- B. Attitudes, interests, and appreciations
 - Civic consciousness (civic pride)
 - Appreciation and understanding of beauty in modern products of all kinds
 - Interest in art museums, travel, and further study
 - Interest in the civic, domestic, and social service of art
- C. Mental technique
 - Good taste, discriminating judgment, ability to select and choose wisely
 - Creative ability, originality, initiative, imagination, keen observation
 - Ability to analyze works of art and to understand the factors of beauty in production
 - Keener observation, beauty of nature and fine things of art
- D. Right habits and skills
 - Constructive thinking and planning
 - Systematic organization
 - Practical technique
 - Co-ordination of mind, hand, and eye
 - Freedom and spontaneity
 - Order, neatness
 - Body and mind training
 - Self-activity
 - Worthy use of leisure time

¹⁵ Adapted from W. G. Whitford, *An Introduction to Art Education*. The Century Co., New York, 1929.

this course is organized around four groups of outcomes which are quite similar to the artistic experiences presented in the previous paragraph.

V. MEASUREMENT OF ART ABILITIES AND ACHIEVEMENT

Three types of tests may be distinguished in the field of art education: (1) drawing scales and tests, (2) art appreciation tests, and (3) art abilities tests. As many of the art tests cannot be illustrated easily, brief descriptions of a few representative scales and tests and occasional illustrations are given on the following pages to familiarize the student with measurement devices in this field.

Drawing Scales and Tests. Several rating scales for use in the evaluation of art achievement are now available. Such scales must depend, as do all scales, upon the representative nature of the specimens selected for presentation and the skill of judges in using the scales. Evidence from a study of the values of drawing scales indicates that their use reduces the inaccuracy of ratings to about one-half of that obtained when no scale is used.¹⁶

The *Kline-Carey Drawing Scales* consist of series of samples for measuring (1) representation, and (2) design and composition. The first series uses such subject matter as a house, a tree in silhouette, a **running boy**, and a **rabbit** in scales having 14 samples, while the second uses the themes of illustrations, posters, structural designs, and borders.

Tests of Art Appreciation. The increasing stress placed upon the appreciative outcomes of art instruction of recent years results in a significant place for art appreciation tests among evaluative tools. Two tests of art judgment or talent are worthy of brief discussion here—the *Meier Art Judgment Test* and the *McAdory Art Test*.

In the *Meier Art Judgment Test*, which may be given as an individual or as a group test, the pupil is confronted with 100 pairs of artistic specimens adapted from many sources. One of each pair of specimens has been changed in some specific element from the original form. The exact feature

¹⁶ Fowler D Brooks, "The Relative Accuracy of Ratings Assigned with and without the Use of Drawing Scales," *School and Society*, 27 518-20; April 23, 1926.

changed is specified in the record sheet on which the pupil records his reaction. A consideration of the complete series of paired specimens insures a comprehensive sampling of the various elements which enter into æsthetic judgment. According to the evidence obtained by the author, this test measures the sensitivity of the individual to the effect which the composition as a whole produces on the observer. In order to give a better idea of the exact nature of specimens and the accompanying record sheets, a single pair of the etchings is reproduced here along with a brief sampling of seven items from the Test Record Sheet. The pair of specimens reproduced here is used with item 49 in the record sheet. In this item, the presence or absence of horns is the point for special consideration in making the judgment. The scoring key lists the drawing with horns as the one of greater merit.

The *Meier Art Judgment Test* supersedes the *Meier-Seashore Art Judgment Test*, and is made up of a smaller number of carefully selected items. This test is the first of a series of three tests to be known as the *Meier Art Tests*. The series when completed will consist of: Part I, Art Judgment, Part II, Creative Imagination; and Part III, Aesthetic Perception. Part II is in preparation and Part III is projected.

The *McAdory Art Test* is somewhat similar to the *Meier Art Judgment Test*, but has only 72 pairs of plates, 24 of which are in color, and calls for reactions to a wide variety of materials, such as furniture, clothing, and rugs.

Art Abilities Tests. Two tests which purport to measure art abilities, mainly the outcomes of art instruction, are the *Lewerenz Test in Fundamental Abilities of Visual Arts* and the *Knauber Test of Art Ability*. Their major values seem to be at the junior and senior high school levels, although the first is designed for use as low as the third grade.

The *Lewerenz Test in Fundamental Abilities of Visual Arts* is designed to measure aspects of art ability in nine areas: (1) recognition of proportion, (2) originality of line drawings, (3) observation of light and shade, (4) knowledge of subject-matter vocabulary, (5) visual memory of proportion, (6) analysis of problems in cylindrical perspective, (7)

EXCERPTS FROM MEIER ART TESTS, I, ART JUDGMENT¹

DIRECTIONS

In the accompanying booklet are pictures arranged in pairs, the two in each pair being very nearly alike. They differ only in *one* respect and you are told *what* that is in each case on pages 1, 2, and 3 of this blank.

You are to compare the two pictures in each pair, noting the unlike portion, and then decide which one is better (more pleasing, more artistic, more satisfying). Do not hurry. Study each pair carefully in turn.

Indicate your preference by making an X in the circle under *Left*, if you decide that the left-hand picture is better, or in the circle under *Right* if you believe that the right hand one is more desirable.

Examples of proper marking (pictures not illustrated)

- | | | | |
|----------------------------------|----------------------------------|----|----------------------------------------------------------------------------------------|
| Left | Right | No | |
| <input checked="" type="radio"/> | <input type="radio"/> | | A Presence or absence of tree. (This would mean that you prefer the left-hand picture) |
| <input type="radio"/> | <input checked="" type="radio"/> | | B Treatment of waves. (This would mean that you prefer the right-hand picture) |

Select the better one in every pair. Do not omit any. If unable to decide within a reasonable time mark the place and return to that one later.

Left	Right	Pair No.	Difference
<input type="radio"/>	<input type="radio"/>	1	Arrangement of wall and foreground
<input type="radio"/>	<input type="radio"/>	2	Foreground
<input type="radio"/>	<input type="radio"/>	49	Inclusion or omission of the horns
<input type="radio"/>	<input type="radio"/>	50	Arrangement in picture of the woman and umbrella
<input type="radio"/>	<input type="radio"/>	51	Position of the figures
<input type="radio"/>	<input type="radio"/>	99	Direction of pine tree's main branch
<input type="radio"/>	<input type="radio"/>	100	Treatment of the water



49

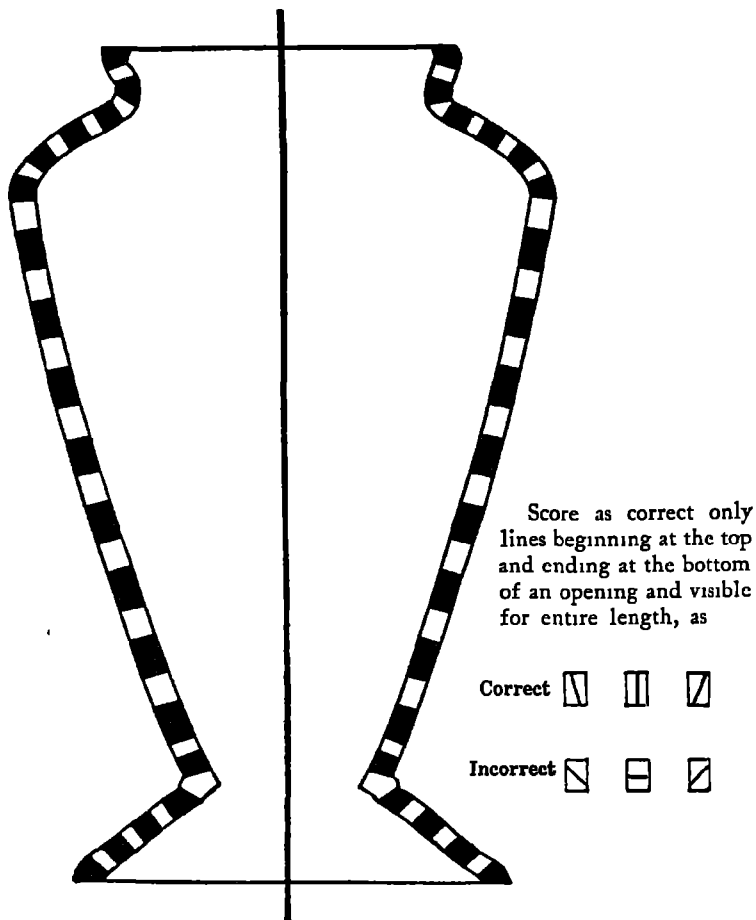
analysis of problems in parallel perspective, (8) analysis of problems in angular perspective, and (9) recognition of color.

One of the most interesting scoring schemes the authors have seen is supplied for use with Test 5. Because of the unusual methods of objectification employed, a copy of the

¹⁷ Norman C. Meier, *The Meier Art Tests, I, Art Judgment*. Published by Bureau of Educational Research and Service, University of Iowa, 1940.

SAMPLE SCORING KEY FOR LEWERENZ TEST IN FUNDAMENTAL
ABILITIES OF VISUAL ARTS¹⁸

TEST 5 VISUAL MEMORY OF PROPORTION



Directions for Scoring Place the scoring key over the paper to be scored so that the intersecting cross lines exactly coincide with those on the paper. The score is determined by the number of openings under which the outline drawn by the student falls.

¹⁸ Alfred S. Lewerenz, *Test in Fundamental Abilities of Visual Arts Scoring Key*. Published by Southern California School Book Depository, 1927.

key used in scoring a drawing of a vase is reproduced on page 458.

Applied Art. Education will have failed in much of its social responsibility if it allows the child to leave the school without developing in him a rather definite love of the fine arts, even though it may be on a relatively low level. Not everyone can or should become a musician, a painter, or a sculptor, but almost everyone has the essential elements which make for a love and appreciation for the beautiful which he himself may be unable to produce. Instruction in the fine arts should cultivate and develop these elements. Furthermore, such instruction has a rather definite responsibility for making the arts function in real life in matters of personal adornment and in the planning and decorating of the home. The general cultural level of the population will be raised as this point is recognized and applied in instruction in the fine arts.

TOPICS FOR DISCUSSION

1. In your opinion is there any reason to assume that achievement in the fine arts may not be objectively measured? Support your answer
2. Is the measurement of musical talent or artistic talent any more difficult or basic than the measurement of aptitudes in any other complex field?
3. What are the major types of aims in music education? •
4. Which of the aptitude tests discussed here seem to be most soundly grounded in critical research?
5. Briefly discuss and illustrate the manner in which musical knowledge and musical skills are measured
6. What is the status of standardized tests of musical appreciation?
7. What similarities in the basic problems of measurement do you see in the fields of music and art?
8. What are the major classes of general outcomes in art instruction?
9. Which of the art tests described here seem most adequately to measure the major features of accomplishment in art?
10. Distinguish clearly between art appreciation and art abilities tests with respect to their functions and forms

SELECTED REFERENCES

- Brooks, Fowler D, "The Relative Accuracy of Ratings Assigned with and without the Use of Drawing Scales" *School and Society*, 27 518-20, April 28, 1928

- Broom, M. E., *Educational Measurements in the Elementary School*, pp. 236-44. New York McGraw-Hill Book Co., Inc., 1939.
- Burns, S. F., "The Value of Prognostic Tests for Instrumental Pupils" *School Music*, 31 6-9, March-April 1931.
- Church, Esther, "The Use of Tests and Measurements in Grouping Music Students of the Junior High School." *Music Supervisors Journal*, 16 79ff, December 1929.
- Dean, Charles D., "Predicting Sight-Singing Ability in Teacher-Education" *Journal of Educational Psychology*, 28 601-8, November 1937.
- Doig, Dorothea, "Can Everyone Use Music Tests?" *Music Educators Journal*, 24 29ff, February 1938.
- Faulkner, Ray, "Standards of Value in Art" *Art in American Life and Education*. Fortieth Yearbook of the National Society for the Study of Education, Chapter XXVII, pp. 401-26. Bloomington, Ill. Public School Publishing Co., 1941.
- Gildersleeve, Glenn, "Standards and the Evaluation and Measurement of Achievement in Music" *Music Education* Thirty-Fifth Yearbook of the National Society for the Study of Education, Part II, Chapter XIX, pp. 195-206. Bloomington, Ill. Public School Publishing Co., 1936.
- Gilliland, A. R., Jordan, R. H., and Freeman, Frank S., *Educational Measurements and the Class-Room Teacher* (Revised Edition), pp. 246-59. New York The Century Co., 1931.
- Graves, Maitland, "What is Your IQ in Design?" *Art Instruction*, 3, 11-14, April 1939.
- Grimes, James W., and Bordin, Edward, "A Proposed Technique for Certain Evaluations in Art." *Educational Research Bulletin*, 18 1ff, January 4, 1939.
- Kerr, Mary R., "Report on Corrective Treatment of a Group of Monotones." *Elementary School Journal*, 23 197-202, 283-95, November and December 1922.
- Kinter, Madaline, *The Measurement of Artistic Abilities*. New York The Psychological Corporation, 1933.
- Klar, Walter H., Winslow, Leon L., and Kirby, C. Valentine, *Art Education in Principle and Practice*. Springfield, Mass. Milton Bradley Co., 1933.
- Kwalwasser, Jacob, "From the Realm of Guess into the Realm of Reasonable Certainty." *Music Educators Journal*, 24 16-17; February 1938.
- Kwalwasser, Jacob, *Tests and Measurements in Music*. Boston C. C. Birchard and Co., 1927.
- Larson, William S., "Practical Experience with Music Tests." *Music Educators Journal*, 24 31ff, March 1938.
- Meier, Norman C., "Diagnosis in Art" *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XXII, pp. 463-76. Bloomington, Ill. Public School Publishing Co., 1935.

- Meier, Norman C, "Recent Research in the Psychology of Art" *Art in American Life and Education* Fortieth Yearbook of the National Society for the Study of Education, Chapter XXVI, pp. 379-400. Bloomington, Ill Public School Publishing Co, 1941.
- Moore, J E., "Art Education." *Encyclopedia of Educational Research*, pp. 58-65. New York The Macmillan Co, 1941.
- Munro, Thomas (Chairman), *Art in American Life and Education*. Fortieth Yearbook of the National Society for the Study of Education. Bloomington, Ill Public School Publishing Co., 1941.
- Nelson, M J., *Tests and Measurements in Elementary Education*, Chapter VIII New York The Cordon Co, 1939.
- Riemenschneider, Albert, "Education of the Music Student Comprehensive Examinations." *Proceedings of the Music Teachers National Association*, pp. 26-32. Oberlin, Ohio Music Teachers National Association, 1935.
- Schoen, Max, "Report of the Committee on Music Tests and Measurements" *Proceedings of the Music Teachers National Association*, pp. 320-50. Oberlin, Ohio Music Teachers National Association, 1935
- Seashore, Carl E, "The Discovery and Guidance of Musical Talent." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XXI, pp. 447-61. Bloomington, Ill Public School Publishing Co, 1935.
- Seashore, Carl E, "Educational Guidance in Music." *School and Society*, 45 385-93, March 20, 1937
- Stanton, Hazel M., *Prognosis of Musical Achievement*. Studies in Psychology, Vol. I, No. 4. Rochester, N. Y. Eastman School of Music, University of Rochester, 1929
- Stanton, Hazel M, *Psychological Tests of Musical Talent*. Rochester, N Y Eastman School of Music, University of Rochester, 1929.
- Todd, Jessie, "A Test in Art for Grade Children." *School Arts Magazine*, 30 365-68, February 1931.
- Uhl, Willis L, "Contributions of Research to Special Methods Music and Art" *The Scientific Movement in Education* Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II, Chapter XIV, pp 171-78 Bloomington, Ill. Public School Publishing Co, 1938
- Uhl, Willis L (Chairman), *Music Education* Thirty-Fifth Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill. Public School Publishing Co, 1936.
- Whitford, William G, *An Introduction to Art Education*. New York: The Century Co, 1929
- Webb, L. W, and Shorwell, Anna Markt, *Testing in the Elementary School*, Chapters XVII-XVIII. New York. Farrar and Rinehart, Inc., 1939.

CHAPTER XX

MEASUREMENT IN HEALTH AND PHYSICAL EDUCATION

This chapter presents a brief summary of the following aspects of measurement and remediation in health and physical education.

- a.* Status and aims of health education.
- b.* Measurement and evaluation in health education.
- c.* Diagnosis and preventive measures for the improvement of health
- d.* Philosophy and objectives of physical education.
- e.* Measurement in physical education.
- f.* Diagnosis of physical condition.

The rather closely related fields of health and physical education are exceedingly important in the school, although they seem not to be as much favored in the curricular setup in most schools as are the academic areas. The national and economic importance of good health is obvious, but the loss to society from illness and unnecessary death is not so apparent as the loss to the individual. Physical education perhaps occupies a more important place in the society of today than was true earlier in the history of man because of the modern need for physical activities to counteract the effects of the physically inactive lives led by many persons.

I. THE SCOPE AND AIMS OF HEALTH EDUCATION

Considered in its broadest sense, health education includes much more than can be treated in this chapter. The mental health, or mental hygiene, aspect is considered in Chapter XI. Some of the health education activities, such as health service, mental hygiene of the classroom, and recreation, are not measurement problems as much as they are supervisory or administrative problems. This chapter deals primarily with health education and physical education measurement, with some attention to the aims and objectives of these fields. Although they are in the main treated separately here, health

and physical education are not equally inclusive terms. Instead, the latter may be considered as one aspect of the former.

Scope of Health Education. Strang states the scope of health education as follows:

Health education may be defined as all the physical conditions, experiences, information, and counsel in and outside of school which produce desirable changes in personal, racial, and community health. Thus defined, health education is clearly recognized as a continuous series of learning experiences built around the whole life of the child.¹

This statement makes clear the great responsibility of the school for the health education of the child, especially because the child not only undergoes physical as well as intellectual experiences of rather broad scope in his school activities but also because his school actions and personality are influenced by his physical activities outside of and beyond the control of the school.

Aims of Health Education. The aims stated by a joint committee of educators and physicians² indicate the purposes underlying health education.

1. To instruct children and youth so that they may conserve and improve their own health

2. To establish in them the habits and principles of living which throu their school life and in later years will aid in providing that abundant vigor and vitality which are a foundation for the greatest possible happiness and service in personal, family, and community life *

3. To promote satisfactory habits and attitudes among parents and adults thru parent and adult education and thru the health education program for children, so that the school may become an effective agency for the advancement of the social aspects of health education in the family and in the community as well as in the school itself

4. To improve the individual and community life of the future, to insure a better second generation, and a still better third generation, to build a healthier and fitter nation and race

Comparatively minor differences in topics comprising the health curriculum are found from school to school in the elementary grades. The major areas of study have to do with: (1) such health habits as cleanliness, food and nutri-

¹ Ruth Strang, "Health Education" *Encyclopedia of Educational Research*, p 560 The Macmillan Co, New York, 1941

² *Health Education* Report of Joint Committee on Health Problems in Education of the National Education Association and the American Medical Association (Second Revision), p 15 National Education Association, Washington, D C, 1941.

tion, sleep and rest, posture and exercise, dental hygiene, ventilation; clothing, first aid and safety; and the use of alcohol and narcotics, and (2) attitudes of courage, helpfulness, consideration of others, independence, adaptability, and enjoyment of daily living.³

II. MEASUREMENT AND EVALUATION IN HEALTH EDUCATION

Classroom testing in the field of health education by the use of paper-and-pencil tests has not attained any significant state of development. Standardized health tests are comparatively new, for it has been almost entirely since 1930 that attention of makers of standardized tests has been directed toward health education. The discussion of health tests to follow is therefore necessarily brief. Some of the measuring instruments discussed in the latter part of this chapter under the heading of physical education have at least indirect significance as health measures.

Health Knowledge Tests. Several good health knowledge tests have recently been published, but most of the tests published prior to 1930 doubtless have little significance today because of the tremendous advances which have been made in nutrition since their publication and the importance of correct knowledge concerning nutrition as a basis for making dietary decisions.

Strang points out that norms are of no great significance for health knowledge tests because the discovery of individual pupil variations and their meaning is much more fundamentally important than are comparisons of individual or group test performance with a norm.⁴

Sample items from two health knowledge tests are given in accompanying illustrations as representative of the testing technique and content of modern tests in this field. One of the best known tests of this type is the *Gates-Strang Health Knowledge Test*, which is a revision of an earlier test by the same authors.

³ Bernice E. Leary, *A Survey of Courses of Study and Other Curriculum Materials Published Since 1934*. U. S. Office of Education Bulletin, 1937, No. 31. Government Printing Office, Washington, D. C., 1937.

⁴ Strang, op. cit. 568.

EXCERPTS FROM GATES-STRANG HEALTH KNOWLEDGE TEST⁵

1. Of these five foods, the most important one for children is
 - a. Meat _____a
 - b. Butter _____b
 - c. Fish _____c
 - d. Sugar. _____d
 - e. Milk. _____e
12. Automobile accidents and hard falls are not likely to happen to you when you
 - a. Are in a great hurry _____a
 - b. Take dares _____b
 - c. Are very tired and sleepy _____c
 - d. Are worried about things at home or school _____d
 - e. Look where you are going and keep your wits about you _____e
40. You use up the most calories when you are
 - a. Asleep in bed _____a
 - b. Lying in bed awake. _____b
 - c. Sitting still. _____c
 - d. Standing. _____d
 - e. Running _____e
50. When a person is in good health, his heart
 - a. Beats fast or slowly according to the needs of the body as a whole _____a
 - b. Always beats at the same rate _____b
 - c. Never skips a beat. _____c
 - d. Always sends the same amount of blood to each part of the body _____d
 - e. Beats more slowly when a person walks or runs than when he sits still. _____e

EXCERPTS FROM NATIONAL ACHIEVEMENT TESTS, HEALTH⁶

1. John cleans his teeth once a week, and goes to the dentist every month. Henry cleans his teeth once a day, and goes to the dentist only when he has a toothache. Frank cleans his teeth twice a day, and goes to the dentist every six months. WHO DOES THE BEST THINGS?
 - f. John
 - t. Henry
 - b. Frank
15. Germ diseases may be caused by
 - a. eating many apples
 - b. drinking pasteurized milk
 - c. lack of vitamin A
 - d. streptococcus serum
 - e. using someone else's tooth-brush
18. Intestinal poisons may cause
 - f. toothache
 - g. giantism
 - b. sore throat
 - x. beriberi
 - z. antiseptics

⁵ Arthur I. Gates and Ruth Strang, *Gates-Strang Health Knowledge Test*, Grades 3 to 8. Published by Bureau of Publications, Teachers College, Columbia University, 1937.

⁶ Robert K. Speer and Samuel Smith, *National Achievement Tests: Health*, Grades 3 to 8. Published by Acorn Publishing Co., 1938.

Health Attitudes Inventories. Comparatively little effective work has been done in the measurement of health attitudes of pupils. Health attitudes can be approached from two standpoints: (1) pupil attitudes toward health practices and beliefs in certain courses of action, and (2) pupil likes and dislikes for various types of foods, activities, and health practices. The same weakness is inherent in these instruments as in attitudes scales in general—there is little evidence to support the belief that expressed attitudes necessarily are borne out in terms of conduct.

The *Health Awareness Test*, from which a few sample items are shown below, is one of the recent publications of this type. It is the result of research in the health measurement field carried on by the American Child Health Association.

EXCERPTS FROM HEALTH AWARENESS TEST⁷

(Directions for matching test given on blackboard and orally.)

- | | | |
|-----------------------------------------|-----|--------------|
| 1. Keep from breeding | () | Wet feet |
| 2. Should not touch other people's food | () | Bad cold |
| | () | Bedroom |
| 3. Blow the nose gently, not hard | () | Garbage pail |
| 4. Keep covered | () | Flies |
| 5. Scald with boiling water | () | Sore throat |
| 6. Should be very clean | () | Babies' milk |
| 7. Should not be too warm | () | Sick people |
| | () | Whiskey |
| | () | Dirty dishes |

DIRECTIONS *If statement number 1 is true, put a circle around the T, but if it is false, put a circle around the F. Do the same for all the statements.*

- | | | |
|------------------------------------------------------------------------------------------------------------------------------------------|---|---|
| 1. Candy should be eaten only at the end of a meal. | T | F |
| 2. An orange, a glass of milk, and hot cooked whole wheat cereal is a better breakfast than an orange, a glass of milk, and puffed wheat | T | F |
| 3. Hot cinnamon rolls or white rolls, fresh and hot, are the best kind of bread for boys and girls. | T | F |

⁷Raymond Franzen, Mayhew Derryberry, and W. A. McCall, *Health Awareness Test*. Published by Bureau of Publications, Teachers College, Columbia University, 1933.

Physical Examinations. Physical examinations in connection with the evaluation of pupil health will be mentioned only briefly here, for they obviously are not the province of the classroom teacher. Health defects often come to light in these examinations, although the annual physical check-ups in some schools may be so perfunctory as to overlook serious health defects.

III. PREVENTION AND DIAGNOSIS IN HEALTH EDUCATION

Diagnosis in health education perhaps more than in any other instructional field must be considered both from the standpoint of school diagnosis and from the usual standpoint of individual pupil diagnosis. Class diagnosis has been discussed in several chapters of this volume, but school diagnosis is here mentioned for the first time.

Diagnosis in the wide range of pupil characteristics which can be considered as related to health is a problem of such constant significance that objective tests can be expected to contribute only in a relatively indirect manner. They cannot be of value in diagnosing contagious diseases and other health conditions which demand immediate attention when cases are found. The diagnostic significance of results from health knowledge tests and health attitudes inventories is not specific. The former are survey rather than specifically diagnostic tools and the latter suffer from the fact that their results are not necessarily indicative of health behavior. Therefore, the major diagnostic possibilities in the field of health education are to be found elsewhere than in the standardized test.

The physical examination has diagnostic significance, of course, but the advantages of continuous measurement and diagnosis are lost when such examinations are made only at infrequent intervals. It is possible, however, for the teacher to supplement the physical examination through his opportunities for the daily observation of pupils in the school. The teacher's place as a diagnostician, non-technical though his diagnoses may be, is fundamentally important. The teacher should recognize that sore throat, vomiting, skin

rashes, and various evidences of contagious colds frequently indicate the desirability of an immediate dispatch of the pupil to the school nurse or to his home. It is through the abilities of teachers to diagnose illness, although not necessarily its specific nature, that individual and group pupil health are protected. The opportunity for such diagnoses of pupil health is usually provided by the morning inspection.

Diagnosis by the teacher can also be made for less immediately important health conditions as the result of continuous observation of pupils. For example, visual defects may be recognized through postural conditions during reading; auditory defects are sometimes major causes of poor spelling, malnutrition frequently results in physical abnormality, goiter of some types is evidenced in a swelling of the thyroid gland in the throat; and such nervous ailments as epilepsy and chorea furnish unmistakable signs. The teacher can often supplement the work of the health agencies of the school by constant alertness for such signs of health defects and by consulting with qualified authorities or referring the case to the proper agencies when he recognizes characteristics he believes to be symptomatic of health defects needing remediation.

Prevention as a phase of diagnosis is clearly important in health education. Isolation of pupils with contagious diseases and immunization of pupils against smallpox, diphtheria, and typhoid fever are preventive responsibilities now accepted by the schools in many communities. Provision of healthful school conditions and a desirable type of school atmosphere and morale are also important as preventive measures.

IV. THE OBJECTIVES OF PHYSICAL EDUCATION

Physical education has come during the past decade increasingly to be thought of as making a valuable contribution to the education process, and its philosophy has consequently been dominated recently by broader aims than were generally held previously. The college and secondary school have better organized programs than do the elementary schools, as less attention has been devoted to physical education for

elementary school children than for high school and college students. The statement of aims which follows represents the modern philosophy concerning the contribution physical education should make to the attainment of desirable educational outcomes in the pupil.

The general objectives of physical education listed by LaPorte⁶ indicate the types of pupil outcomes to which a good physical education program should lead.

1. The development of fundamental skills in aquatic, gymnastic, rhythmic, and athletic activities for immediate educational purposes—physical, mental, and social.
2. The development of useful and desirable skills in activities suitable as vocational interests for use during leisure time
3. The development of essential safety skills and the ability to handle the body skillfully in a variety of situations for the protection of self and of others.
4. The development of a comprehensive knowledge of rules, techniques and strategies in the above activities suitably adapted to various age levels.
5. The development of acceptable social standards, appreciations and attitudes as the result of intensive participation in these activities in a good environment and under capable and inspired leadership
6. The development of powers of observation, analysis, judgment, and decision through the medium of complex physical situations
7. The development of the power of self-expression and reasonable self-confidence (physical and mental poise), by mastery of difficult physical-mental-social problems in supervised activities.
8. The development of leadership capacity by having each student within the limits of his ability, assume actual responsibility for certain activities under careful supervision
9. The elimination of remediable defects and the improvement of postural mechanics insofar as these can be influenced by muscular activities and health advice, based on adequate physical and health diagnosis.
10. The development of essential health habits, health knowledge and health attitudes as the result of specific instruction in health principles and careful supervision of health situations.

V. MEASUREMENT IN PHYSICAL EDUCATION

Persons interested in testing and evaluating various aspects of physical ability and skill will find almost no commercially published paper-and-pencil tests for those purposes. On the

⁶ William R. LaPorte, "The Ten Major Objectives of Health and Physical Education" *California Physical Education, Health and Recreation Journal*, p. 6 January, 1936

other hand, they will find voluminous reports of testing and evaluative techniques in the educational literature, both books and journals. It should be apparent that measures of physical fitness, motor ability, and physical skills must be conducted by means of physical and medical measurements and tests and by observation of behavior in situations involving physical activity rather than by the use of standardized tests.

Tests of General Physical Qualities. Tests of such qualities, commonly thought to be inherited, as motor ability, physical capacity, and athletic ability are considered under this heading. Each of these tests consists in the main of various measures of motor abilities and physical skills which are combined into a composite score. Their results are useful variously as a basis for classifying pupils into groups for physical education and for predicting levels of physical attainment.

Rogers has devised a series of physical tests for administration to individual pupils from which two types of derived scores are obtained—a strength index and a personal fitness index.⁹ The tests, having different procedures in some instances for boys and girls, are accompanied by tables of normal strength indices differentiated for age and sex groups. Their major purpose is to determine deficiencies and to facilitate classification of pupils into groups having common remedial needs. No presentation of these tests can be given here because of their detailed nature.

A test of general motor capacity based on various specific tests of track and field events and of strength was developed by McCloy.¹⁰ Something approaching a profile of the individual's general capacity is furnished by the results of this test, which has particular value in the prediction of ultimate levels of attainment.

The *Iowa Revision of the Brace Test of Motor Ability* consists of 21 physical stunts yielding a composite score of

⁹ Frederick R. Rogers, *Physical Capacity Tests* A. S. Barnes and Co., New York, 1938.

¹⁰ C. H. McCloy, "The Measurement of General Motor Capacity and General Motor Ability" *Research Quarterly of the American Physical Education Association*, 5 46-61, March 1934.

motor educability.¹¹ The tests are so devised and set up that pupils can do the scoring and the recommended procedure is that one-half of the class score the other half on performance of the stunts and that the two groups then be reversed. Scores are interpreted in terms of T-score values.

Johnson devised a test of motor educability which has values for the sectioning of classes.¹² Although it is slower and more difficult of administration than the *Iowa-Brace* test it is rated as probably the best available test of motor educability.

Cardiovascular Tests. Good and poor physical condition can be determined by cardiovascular tests involving pulse rate and blood pressure under varying conditions of rest and fatigue. Several of the tests of this type based on pulse counts are sufficiently easy to administer and require so little equipment that they are subject to use by the skilled teacher.¹³ The significance of such tests is somewhat reduced by the fact that most of them measure only one type of physiological efficiency, whereas some other of the important variables of blood pressure, pulse rate, and related functions are not well enough understood at this time to be included in these tests.

Posture Tests. Posture tests cannot be administered in a routine manner in the usual school situation because of their complex nature. Most of the tests of posture are based on comparisons of pupil silhouettes with silhouettes representing standard posture or representing several degrees of postural merit from very poor to excellent. Because of the wide variability in posture and the incomplete nature of evidence on the question, Cozens suggests that teachers should probably not attempt to make pupils conform to any pattern considered desirable.¹⁴

¹¹ C H McCloy, "An Analytical Study of the Stunt Type Test as a Measure of Motor Educability" *Research Quarterly of the American Physical Education Association*, 8 46-55, October 1937

¹² Granville B Johnson, "Physical Skill Tests for Sectioning Classes into Homogeneous Units" *Research Quarterly of the American Physical Education Association*, 3 128-36, March 1932

¹³ Charles H McCloy, *Tests and Measurements in Health and Physical Education*, Chapter XX F S Crofts and Co, New York, 1939

¹⁴ Frederick W Cozens, "Physical Education — Measurement" *Encyclopedia of Educational Research*, p 815 The Macmillan Co, New York, 1941

General Achievement Scales. General achievement scales have been developed for the measurement of ability in various sports activities. Their purposes are to stimulate pupil interest and performance, to determine the sports skills of individual pupils and groups, and to diagnose deficiencies. Such scales are highly time consuming, however, and have not been adequately validated.¹⁵

Knowledge and Information Tests. Paper-and-pencil tests of knowledge and information in specific sports activities and comprehensively for all activities have appeared in the physical education journals but have not been published commercially in standardized form. The following illustrations indicate the manner in which various objective item forms are adaptable to measurement of knowledge and information in this field.¹⁶

TRUE-FALSE ITEMS

Encircle the correct answer.

- T F The follow-through in a golf drive determines the accuracy of the flight of the ball.
- T ? F There are African negro tribes who have athletes able to high jump to heights greater than the present American record

Encircle the correct answer. If the answer is false, cross out the word which makes it false and insert the word that makes it true.

- anopheles
- T F The ~~culex~~ mosquito is the transmitter of the malaria germ.

Underline T if the statement is true and F if the statement is false. If the converse of the statement is true, underline CT, if the converse is false, underline CF

- T F CT CF Low arches are always painful

MULTIPLE-CHOICE ITEMS

Place an X in the space before the phrase which correctly completes the statement.

¹⁵ Inasmuch as these scales are too highly specialized to warrant presentation, or even illustration, here, the student should refer to the bibliography at the end of this chapter for source materials.

¹⁶ McCloy, op. cit pp. 190-97.

The world's record for the mile run is approximately:

- _____ 3' 20"
- _____ 4' 6"
- _____ 8' 11"
- _____ 2' 19"

Check the one or more correct answers under each statement.

According to the currently accepted "best" form for the shot-put (for a right-handed putter):

- _____ In the hop, the right foot alights well before the left foot.
- _____ The shot should remain as close to the neck as possible.
- _____ The reverse is of no importance, and is just a traditional movement.
- _____ The shot should be held deep in the palm of the right hand.
- _____ The best angle (to the ground) of the putting effort is approximately forty-one degrees.

MATCHING EXERCISES

In the following questions, write the number belonging to the approximately correct date in the first space and the number corresponding to the correct name in the second space.

At about the year

- _____, a physical education program was introduced at the Philanthropium in Dessau by _____.
- _____, physical education was established at the Round Hill School in the United States by _____.
- _____, a department of physical education was opened in the Y M. C. A Training School at Springfield, Massachusetts, under the guidance of _____.
- _____, the King of Denmark appointed as professor of physical education in the university _____.
- _____, the modern Olympic Games were revived, largely because of the work of _____.

Dates

Names

- | | | | |
|---------|----------|-----------------|---------------|
| 1. 1776 | 6. 1887 | 1. Basedow | 6. Hitchcock |
| 2. 1799 | 7. 1897 | 2. Beck | 7. Jahn |
| 3. 1804 | 8. 1902 | 3. Bukh | 8. Ling |
| 4. 1810 | 9. 1906 | 4. de Coubertin | 9. McCurdy |
| 5. 1823 | 10. 1924 | 5. Gulick | 10. Nachtgall |

COMPLETION EXERCISES

Fill in the blank spaces with the words which most accurately complete the statement.

In the high school low hurdles race, the distance from the start to the first hurdle is _____ yards, it is _____ yards between the hurdles, and it is _____ yards from the last hurdle to the finish. There are _____ hurdles to be cleared.

Tests of Proficiency in Sports. Numerous articles in the physical education journals present tests of technique in a variety of sports.¹⁷ These tests are usually based upon an analysis of the skills involved in the sport and the construction of tests for the measurement of those skills. Validation of the batteries of tests is by means of comparisons between scores made by pupils and teachers' judgments of pupil proficiency. Cozens lists the *Heath-Rodgers Soccer Test for Elementary School Boys*, the *Dyer Backboard Test of Tennis Ability*, and the *French-Cooper Volleyball Test for High School Girls* as those at the elementary and junior high school levels which have sufficient validity for other than group comparisons.¹⁸

Physical Classification Tests. The importance of tools to be used in the classification of pupils for physical education and particularly for competitive sports is obvious. Physical differences among pupils of the same age are so great that classification by chronological age is likely to result in injuries to the smaller and weaker children and usually deprives them of adequate opportunities for exercise. Two indices useful for classification purposes at the elementary and junior high school levels have been validated. As the brief indications of their nature make clear, such indices are obtained by the use of physical rather than paper-and-pencil tests.

McCloy developed a classification index for elementary school children.¹⁹ The formula is as follows:

$$\text{Classification Index} = 20A + 6H + W,$$

¹⁷ See bibliography at end of this chapter for such references.

¹⁸ Cozens, *op cit* p 816.

¹⁹ C H McCloy, *The Measurement of Athletic Power*. A S. Barnes and Co., New York, 1932.

where A refers to age in years, H to height in inches, and W to weight in pounds. Another index making use of the same physical and age measures has been derived for junior high school girls.²⁰ The index is obtained by the use of the formula:

$$\text{Index} = 2A + H + .11W.$$

Cozens states that the factors of age, weight, and height when combined in the proper combination probably are nearly as useful for classification purposes as are more complex measures and the simplicity of their use is of considerable importance.²¹

VI. DIAGNOSIS IN PHYSICAL EDUCATION

Diagnosis in physical education as well as in health education appears to depend much more upon teacher observation and physical examinations than upon any standardized testing devices of the pencil-and-paper type. The tests of general physical qualities and of physical fitness serve some diagnostic functions. Other tests of diagnostic value are those for the measurement of blood pressure under varying conditions of fatigue. Both of these types can be given by a skilled teacher. Still other tests require technical knowledge and equipment not ordinarily possessed by the teacher.

A significant trend in diagnosis in physical education is that the issue is being approached from the functional rather than the structural standpoint. Even with functional tests, however, it is felt by some that the tests fail to measure completely enough, particularly with respect to such organs as the nervous system, to furnish a highly satisfactory diagnostic score.²²

TOPICS FOR DISCUSSION

1. Discuss the aims of health education
2. Comment upon the nature and present status of health knowledge testing

²⁰ F W Cozens, Hazel J Cubberley, and N P Neilson, *Achievement Scales in Physical Education Activities for Secondary School Girls and College Women*. A S Barnes and Co., New York, 1937

²¹ Cozens, op cit p 815

²² Whitelaw R Morrison and Laurence B Chenoweth *Normal and Elementary Physical Diagnosis*, pp 331-33 Lea and Febiger, Philadelphia, 1932

3. What is a major limitation of health attitudes inventories?
4. Discuss some of the preventive and diagnostic procedures in health education for use in the classroom.
5. What are the major objectives of physical education?
6. In what way are some of the measures of general physical qualities useful indications of health status?
7. In what way are cardiovascular and posture tests useful in physical education?
8. Illustrate some methods of testing knowledge and information in physical education.
9. Indicate the nature of tests of proficiency in sports
10. Give some of the procedures useful in the classification of pupils for physical education
11. Discuss diagnostic methods in physical education.

SELECTED REFERENCES

- Bovard, John F., and Cozens, Frederick W, *Tests and Measurements in Physical Education* (Second Edition, Revised). Philadelphia W. B. Saunders Co., 1938
- Brace, David K., "The Development of Measures of Pupil Achievement in Physical Education" *Research Quarterly of the American Physical Education Association*, 2 32-37, October, 1931.
- Brace, David K, "Testing Basketball Technique." *American Physical Education Review*, 29 159-65, April 1924.
- Brueckner, Leo J, and Melby, Ernest O, *Diagnostic and Remedial Teaching*, Chapter XIV. Boston. Houghton Mifflin Co., 1931.
- Carver, Margaret, "Motivation of Child Interest in Corrective Physical Education in Elementary Schools." *Journal of Health and Physical Education*, 4 27ff, October 1933.
- Cozens, Frederick W, "A Curve for Devising Scoring Tables in Physical Education." *Research Quarterly of the American Physical Education Association*, 2 67-75, December 1931.
- Cozens, Frederick W, "Physical Education — Measurement." *Encyclopedia of Educational Research*, pp. 814-18. New York: The Macmillan Co, 1941
- Crapser, A. Lester, "National Physical Achievement Standards." *Journal of Health and Physical Education*, 1.14ff.; January 1930.
- Edgren, H. D, "An Experiment in the Testing of Ability and Progress in Basketball" *Research Quarterly of the American Physical Education Association*, 3 159-71, March 1932
- Eginton, Daniel P, "Tests and Measures of Satisfactory Growth." *Educational Method*, 15 145-49, December 1935
- Gudakunst, Don W, "Diagnosis in Health Education." *Educational Diagnosis* Thirty-Fourth Yearbook of the National Society for the Study of Education, Chapter XVII, pp. 347-61. Bloomington, Ill.: Public School Publishing Co., 1935.

- Heath, Marjorie L., and Rodgers, Elizabeth G., "A Study in the Use of Knowledge and Skill Tests in Soccer." *Research Quarterly of the American Physical Education Association*, 3 33-53, December 1932
- Howe, Eugene C., "The Precision and Validation of Tests of Physical Fitness" *Research Quarterly of the American Physical Education Association*, 1 90-96, May 1930.
- Howland, Amy R., "National Physical Education Standards for Girls," *Journal of Health and Physical Education*, 8 223ff., April 1937.
- Kuntz, Alfred, "Therapy for the Maladjusted" *Journal of Health and Physical Education*, 8 143ff., March 1937.
- Lockhart, Aileene, "A Survey of Testing in Tennis" *Journal of Health and Physical Education*, 9 433ff., September 1938.
- Lowman, Charles L., Colestock, Claire, and Cooper, Hazel, *Corrective Physical Education for Groups*. New York A. S. Barnes & Co., 1932.
- McCloy, Charles H., *Appraising Physical Status Methods and Norms*. University of Iowa Studies in Child Welfare, Vol. XV, No. 2. Iowa City University of Iowa, 1938
- McCloy, Charles H., *Appraising Physical Status The Selection of Measurements*. University of Iowa Studies in Child Welfare, Vol. XII, No. 2. Iowa City University of Iowa, 1936
- McCloy, Charles H., *Tests and Measurements in Health and Physical Education*. New York F. S. Crofts and Co., 1939
- Nash, J. B. (Editor), *Interpretations of Physical Education Nature and Scope of Examinations*. New York A. S. Barnes and Co., 1931.
- Neilson, Neils P., and Cozens, Frederick W., *Achievement Scales in Physical Education Activities for Boys and Girls in Elementary and Junior High Schools*. New York A. S. Barnes and Co., 1934
- Palmer, George T., "Measurement of Nutritional Status." *Mind and Body*, 38 452-56, April 1931.
- Rathbone, Josephine L., *Corrective Physical Education*. Philadelphia W. B. Saunders Co., 1934
- Research Staff, American Child Health Association, *Physical Defects: The Pathway to Correction*. New York American Child Health Association, 1934.
- Rodgers, Elizabeth G., and Heath, Marjorie L., "An Experiment in the Use of Knowledge and Skill Tests in Playground Baseball." *Research Quarterly of the American Physical Education Association*, 2 113-31, December 1931
- Rogers, Frederick R., *Physical Capacity Tests*. New York A. S. Barnes and Co., 1938.
- Strang, Ruth, "Health Education" *Encyclopedia of Educational Research*, pp. 561-71. New York The Macmillan Co., 1941.
- Tiegs, Ernest W., *The Management of Learning in the Elementary Schools*, Chapter XV. New York Longmans, Green and Co., 1937.
- Tiegs, Ernest W., *Tests and Measurements for Teachers*, pp. 409-15. Boston Houghton Mifflin Co., 1931.

- Wayman, Agnes, "What to Measure in Physical Education." *Research Quarterly of the American Physical Education Association*, 197-110, May 1930.
- Whitney, Anne, "The Weighing and Measuring of School Children" *Mind and Body*, 38:446-51, April 1931.
- Wilson, Guy M., and Hoke, Kremer, J., *How to Measure* (Revised and Enlarged Edition), Chapter XIV. New York: The Macmillan Co., 1929.

CHAPTER XXI

MEASUREMENT OF GENERAL EDUCATIONAL ACHIEVEMENT

This chapter treats the following points in the measurement of general educational achievement

- a.* Advantages of general achievement batteries.
- b.* Limitations of general measures of achievement.
- c.* General *vs.* specific surveys.
- d.* Types of achievement batteries
- e.* Some distinctive features of certain achievement batteries.

I. GENERAL MEASURES OF ACHIEVEMENT

The emphasis throughout this volume is rather definitely on diagnostic and analytical testing and on evaluative techniques in subject matter and performance areas. However, a consideration of the practical problems of measurement in the classroom leads to the conviction that there is a real service to be rendered by general survey tests of achievement. Accordingly, tests of that type are treated briefly here.

General vs. Specific Diagnosis. The battery-type of general achievement test opens up certain types of possibilities for diagnostic, analytical, and remedial work. Such a test affords a survey of the total instructional situation. It presents a perspective of the measurable aspects of accomplishment. The profile chart, through its valleys and peaks, points out general areas of weakness and strength which need much more detailed and analytical study. A general survey test of any of the types described in this chapter may reveal a specific weakness in, for example, the language skills. To the critical teacher this is a challenge to discover more exactly the factors underlying this limited achievement. Accordingly, the few cases identified by the general achievement test should be subjected to a detailed analytical test in the subject for the purpose of locating specific difficulties and their causes.

It is important that the teacher keep clearly in mind the fact that measurement of general achievement is not a substitute for diagnostic and analytical measurement. It is merely the stepping-stone to it. It cannot be repeated too emphatically that throughout all of the supervisory and instructional uses of tests, measurement which is general and indefinite is likely to be useless. Measurement which points out vague generalities is useful only as a point of departure for more definite measurement. In general, measurement which does not aid in the identification of individual pupil difficulties and their causes is futile, and correction of educational deficiencies without the assistance of exact analytical and diagnostic measures verges on the miraculous.

Advantages and Disadvantages of the Battery-Type Tests of General Achievement. Among the specific qualities of the battery-type tests of general achievement which have been given considerable emphasis by persons interested in the improvement of classroom measurement are the following:

Curricular Content. Many of the shorter tests which are designed to measure in only a single subject-matter field frequently do not closely conform to modern curricular content. The battery test, through its greater length, appears to afford a better basis for critical selection of content. On the other hand, the properly constructed test of a single subject field opens up possibilities for analysis and measurement in the subject which no battery test sampling over a wide range of subject matter can afford.

Units of Measurement. The use of a uniform unit of measurement in the scaling of battery tests constitutes a real advantage in the interpretation of the test results and in the comparisons of results from one subject-matter field to another. While this is an important advantage, it does not at all mean that uniformity in units of measurement may not be secured in single tests in unrelated subjects. As a matter of fact, special scales for the commensuration of the scores from the various parts of battery tests are in common use. Examples of these devices are seen in the *Stanford Achievement Tests* and in a somewhat less convenient form in the *Unit Scales of Attainment*.

Ease of Administration. The tendency of the authors of battery-type tests to utilize the same or similar types of testing techniques throughout the series of tests unquestionably does tend to simplify the problems of administering the test. The use of uniform methods of recording the pupil's responses also simplifies the problem of scoring the tests. In general, however, such battery tests are usually so long that the labor involved in scoring them becomes quite great. One of the very significant criticisms of the *Stanford Achievement Tests*, the *Metropolitan Achievement Examinations*, the *Unit Scales of Attainment*, or any of the other longer batteries is the fact that the time required for administration and for scoring is extremely great. Some experience in the use of such batteries of general accomplishment tests indicates that it requires a speedy clerk to score more than four of these advanced tests in an hour. However, it may be that this is not too high a price to pay for extensive sampling and reliable measurement. Furthermore, one of these tests, the *Stanford Achievement*, is now available for either hand- or machine-scoring.

Simplicity of Interpretation. The use of comparable units of measurement and similar testing techniques in the several tests comprising a general achievement battery simplifies the problems of comparing and interpreting the results. The raw test scores are readily turned into scale values, educational ages, grade equivalents, etc. Modern graphic methods of summarizing test results make effective use of such derived scores. Profile charts of the type used with the *Metropolitan Achievement*, the *Stanford Achievement*, the *Unit Scales of Attainment*, and the *Modern School Achievement* tests add to the clearness with which such test results may be interpreted. Naturally such profiles are useful only in case test scores from a number of different types of measures are reducible to a common unit of measurement.

Unity of Population in Standardization. The fact that the standardization of most comprehensive batteries is based upon returns from the same individual pupils for each of the different subject tests insures a better picture of the relationships of achievement in these different subjects. The

relation of achievement in reading of pupils of a given age or grade to the language achievement of pupils of the same age or grade can be obtained only when tests are standardized under these conditions.

Economy. Any economy in testing which results from the use of battery tests appears to be conditioned by the assumption that measurement is preferably broad and general in its scope and never specific. It is probably true that almost any one of the modern batteries of achievement tests will furnish a wider sampling into more subject-matter fields at a lower cost per pupil than could be accomplished by the selection of single-subject tests for the purpose. There are numerous occasions, however, when it is of greater importance to measure more intensively a limited range of subjects. For this type of measurement the battery tests are clearly not the most economical. In order to provide for this situation, the authors of most test batteries have prepared the tests for certain subjects in separate form. The main difficulty here lies in the fact that the use of any adequate number of these special separate tests runs up the total cost of measurement more rapidly than would result from the careful selection of more comprehensive tests of single subject-matter fields. After all, economy in measurement must be considered in terms of the net cost per unit of valid and reliable information secured. For some purposes the battery test may be most economical, and for other purposes the choice will lie in another direction.

II. TYPES OF GENERAL ACHIEVEMENT BATTERIES

A description of general achievement batteries written ten years ago would, of necessity, have confined its attention to one or two such tests. Now there are many general batteries, several of which have very distinct merit. Only six of the better known and more widely used achievement batteries are described here as types. No attempt is made here to illustrate their measurement techniques, for the extremely wide variety of subject matter tested makes that impracticable. Moreover, numerous illustrations from these tests appear in preceding chapters of this volume.

TABLE XVIII
SUMMARY OF STANFORD ACHIEVEMENT TESTS
Primary Battery

Test No.	Test Name	Number of Items	Testing Technique	Working Time
1	Paragraph Meaning	46	Completion	20
2	Word Meaning ¹	40	Multiple-choice 3	5
3	Spelling	50	Sentence dictation	15
4	Arithmetic Reasoning ²	20	Problems	10
5	Arithmetic Computation	36	Examples	15
¹ Begins second sitting ² Begins third sitting				

Intermediate and Advanced Batteries

Test No	Test Name	Number of Items	Testing Technique	Working Time
1	Paragraph Meaning	45	Completion	20
2	Word Meaning	50	Multiple-choice 5	10
3	Language Usage ¹	100	Alternate-response	15
4	Arithmetic Reasoning	40	Problems	20
5	Arithmetic Computation ²	62-65 ⁴	Examples	30
6	Literature	50	Multiple-choice 3	10
7	Social Studies I ³	50	Multiple-choice 3	10
8	Social Studies II	50	Multiple-choice 3	10
9	Elementary Science	50	Multiple-choice 3	10
10	Spelling	50	Sentence dictation	15
¹ Begins second sitting ² Begins third sitting ³ Begins fourth sitting ⁴ 62 Intermediate; 65 Advanced				

The Stanford Achievement Tests. The original battery of *Stanford Achievement Tests*, designed by Kelley, Ruch, and Terman and copyrighted in 1923, represents one of the outstanding measuring instruments of that period. These tests set new standards of validity, reliability, sampling, test-

ing techniques, standardization, and interpretation for later workers. They undoubtedly did much to stimulate the improvement of measurement in general. After six years, and on the basis of much critical analysis, observation, and experimentation, these tests were revised in the form known as the *New Stanford Achievement Tests*. They have recently been revised a second time, and are again known as the *Stanford Achievement Tests*.

In their present form the *Stanford Achievement Tests* are available in a primary examination for grades two and three, an intermediate examination for grades four to six, and an advanced examination for grades seven to nine. Each examination is available in five forms—D, E, F, G, H. The accompanying tabulations give the names of the tests, the number of items in each, the testing technique used, and the working-time limits separately for the primary and for the intermediate and advanced examinations.

The Cooperative Achievement Tests. The *Cooperative Achievement Tests* for use at the junior high school level consist of tests in three phases of English and reading and in the social studies, natural sciences, and mathematics. These instruments are obtained either in four or in six booklets, depending upon whether the English and reading tests are preferred in the one-booklet form or in separate booklets. Scaled score norms comparable for all tests are furnished. Table XIX shows the nature of these tests, time requirements, and the nature of the abilities measured.

The Iowa Every-Pupil Tests of Basic Skills. These tests of basic skills are provided in four booklets for use during four testing periods. They measure two aspects of the receptive language art in *Tests A* and *B*, one expressive language art in *Test C*, and arithmetic abilities in *Test D*. Consideration of the skills tested by the various parts of the four tests will indicate the fundamentally important skill areas the battery of tests covers.

The Unit Scales of Attainment. The *Unit Scales of Attainment* are a relatively new development in the measurement of general achievement. While the general function of these tests is practically identical with that of the *Stanford Achievement Tests*, the *Metropolitan Achievement*

Tests, etc., the actual methods used in their arrangement are somewhat different. The subjects measured are shown in Table XXI. The multiple-choice type of exercise predom-

TABLE XIX
SUMMARY OF COOPERATIVE ACHIEVEMENT TESTS FOR THE
JUNIOR HIGH SCHOOL

Test Name	Work- ing Time	Abilities Measured (Parts)
English — Mechanics of Expression A .	40	Grammatical Usage Punctuation and Capitalization Spelling
English — Effectiveness of Expression B1 .	40	Sentence Structure and Style Active Vocabulary Organization
English—Reading Comprehension C1	40	Vocabulary Speed of Comprehension Level of Comprehension
Social Studies for Grades 7, 8, and 9	40	Facts, Skills, and Applications Terms and Concepts Comprehension and Interpretation
Science for Grades 7, 8, and 9 ..	40	Facts, Skills, and Applications Terms and Concepts Comprehension and Interpretation
Mathematics for Grades 7, 8, and 9	40	Skills Facts, Terms, and Concepts Applications Appreciations

inates, although the proof-reading or recognition-correction form is used in the English tests.

The distinctive feature of the *Unit Scales of Attainment* is in the arrangement of the scaled items in the divisions of the test. The tests, which are designed for use in grades one

to eight inclusive, are arranged in four divisions—Primary Division for grades one to three, Division 1 for the fourth grade, Division 2 for the fifth and sixth grades, and Division 3 for the seventh and eighth grades. Each division contains items of difficulty suited to the ability of the grades in which

TABLE XX
SUMMARY OF IOWA EVERY-PUPIL TESTS OF BASIC SKILLS,
ELEMENTARY

Test	Title	Part	Skills
A	Silent Reading Comprehension.	I	Reading Comprehension
		II	Vocabulary
B	Work-Study Skills	I	Map Reading
		II	Use of References
		III	Use of Index
		IV	Use of Dictionary
		V	Alphabetization
C	Basic Language Skills	I	Punctuation
		II	Capitalization
		III	Usage
		IV	Spelling
		V	Sentence Sense
D	Basic Arithmetic Skills	I	Vocabulary and Fundamental Knowledge
		II	Fundamental Operations
		III	Problems

it is to be used. This is shown much more clearly by the accompanying analysis of the reading test items. Division 1 contains ten reading comprehension paragraphs representing forty comprehension items. Paragraphs 5 to 10 inclusive are used in Division 2 as Paragraphs 2 to 7. Paragraphs 7 to 10 of Division 1 also appear as paragraphs 1 to 4 in Division 3. Easier items are used at the beginning of Division

1 to afford simple enough stimuli for the lower grade pupils. Paragraphs 7 and 8 of Division 3 are also especially difficult in order to afford adequate top for the advanced grades. The net effect of this scaling process and shifting of items into earlier stages of the test for the advanced grades is to produce a test which contains a larger amount of test material suitable for use in a specific grade than is otherwise possible.

TABLE XXI
SUMMARY OF UNIT SCALES OF ATTAINMENT

Test Name	Primary Division					Division		
	Grade I		Grade II		Grade III	1	2	3
	First Half	Second Half	First Half	Second Half		Grade IV	Grades V VI	Grades VII VIII
Reading—Word Recognition	X	X	X					
Reading—Word Comprehension	X							
Reading—Phrase Comprehension		X						
Reading—Sentence Comprehension			X	X				
Reading—Word Meaning				X	X			
Reading—Paragraph Comprehension					X	X	X	X
Arithmetic Problems					X	X	X	X
Arithmetic Fundamentals					X	X	X	X
Spelling					X	X	X	X
Geography						X	X	X
Literature						X	X	X
Elementary Science						X	X	X
American History						X	X	X
English—Capitalization						X	X	X
English—Punctuation						X	X	X
English—Usage						X	X	X

This plan of shifting test items is not followed exactly in all of the tests in this battery, but the principle holds in most parts. It is nicely illustrated in the section on usage in the English test. The first five items of Division 1 are too easy for any other division. Items 6 to 30 of Division 1 become items 1 to 25 of Division 2. To provide items of adequate

difficulty at the top of Division 2, five new items of higher scale values are added. In turn, Items 6 to 30 of Division 2 appear as Items 1 to 25 of Division 3 with five more items of greater difficulty added to provide top.

TABLE XXII
ARRANGEMENT OF ITEMS IN UNIT SCALES OF ATTAINMENT

Reading Paragraph	Division		
	1	2	3
	Grades 3-4	Grades 5-6	Grades 7-8
1	1		①
2	2	②	②
3	3	③	③
4	4	④	④
5	⑤	⑤	⑤
6	⑥	⑥	⑥
7	⑦	⑦	7
8	⑧	⑧	8
9	⑨	⑨	
10	⑩		

Each of the tests in this battery is standardized on a mental-age basis. Raw test scores in all cases are changed into C-scores by means of specially prepared scales. The accompanying C-score scale represents the C-score equivalents for the raw scores on the English usage section of Division 3.

The Metropolitan Achievement Tests. The *Metropolitan Achievement Tests* represent an ambitious development in this type of measurement. The following list of characteristics of the tests was abstracted from the supervisor's manual:¹

1. This series of tests provides continuous, comparable measures for the entire range of grades from 1 through 8, in all the major elementary school subjects

¹ Richard D. Allen, et al., *Supervisor's Manual Metropolitan Achievement Tests*, pp. 63-75 World Book Co., Yonkers-on-Hudson, N. Y., 1935.

2. The tests are available in the form of complete batteries covering all subjects of the respective grades, as partial batteries covering the core subjects only of the middle and upper grades, and as separate subject tests
3. The division of the series into four batteries, so that none covers more than three grades, makes possible a large number of questions for each grade and therefore thorough measurement.
4. The reliability of the tests is exceptionally high Every single subject test, as well as every battery, accurately measures individual achievement.

TABLE XXIII

C-SCORE EQUIVALENTS FOR ENGLISH USAGE TEST OF
UNIT SCALES OF ATTAINMENT, DIVISION 3

No Right	C-Score	No. Right	C-Score	No Right	C-Score	No. Right	C-Score
0	41	8	70	16	82 5	24	95
1	51	9	72	17	84	25	97
2	56	10	73 5	18	85 5	26	100
3	59	11	75	19	87	27	103
4	62	12	77	20	88 5	28	106
5	65	13	79 5	21	90	29	110
6	67	14	80	22	92	30	115
7	68 5	15	81	23	93 5		

5. The content of the tests is based upon careful checking of many courses of study and the most widely used text-books The questions included were fully validated by expert criticism, try-outs, and rigorous statistical evaluation.
6. The mechanical features of the tests have been perfected as a result of years of experience with test materials Directions for use, keys for scoring, charts for records and for analysis and diagnosis have all been planned to insure best results with a minimum of labor.
7. A special Supervisor's Manual gives complete information for interpreting test results and applying the interpretations for improving school work.
8. Sufficient alternative forms are prepared for retesting from year to year ; there are three forms for the primary grades and five forms for the middle and upper grades.
9. Differential norms have been established so that any school may make comparison with norms based upon conditions most valid for its particular situation.

The Modern School Achievement Tests. The *Modern School Achievement Tests* are similar in most respects to the other batteries previously discussed in this chapter. The accompanying table gives a complete picture of the fields of subject matter measured by these tests.

TABLE XXIV
SUMMARY OF MODERN SCHOOL ACHIEVEMENT TESTS

Test Name	Number of Items	Testing Technique	Working Time
Reading Comprehension	34	Multiple-choice 5 . .	30
Reading Speed and Accuracy	50	Multiple-choice 4	5-8
Arithmetic Computation ¹	35	Examples . .	20
Arithmetic Reasoning	35	Problems	25
Spelling ²	50	Sentence dictation	15
Health Knowledge	56	Multiple-choice 5	15
Language Usage	60	Multiple-choice 3	10
History and Civics ³	60	Multiple-choice 4	15
Geography	60	Multiple-choice 4	15
Elementary Science	50	Multiple-choice 4	12
¹ Begins second sitting ² Begins third sitting ³ Begins fourth sitting			

Other General Achievement Batteries. Not all of the batteries of general achievement tests can be discussed here. The six which are described in this chapter serve merely as illustrations of tests of this type. Among the other general achievement tests which have rendered a real service in the past are the *Progressive School Achievement Tests*, the *Public School Achievement Tests*, the *Illinois Examinations*, the *Otis Classification Test*, *Pintner's Educational Achievement Tests*, and *Lippincott-Chapman Classroom Products Survey Tests*.

TOPICS FOR DISCUSSION

1. In your judgment what is the major supervisory function rendered by the general achievement test batteries?
2. Catalogue the specific advantages presented for batteries such as the

Stanford Achievement Tests, the *Unit Scales of Attainment*, the *Metropolitan Achievement Tests*, and the *Iowa Basic-Skills Tests*.

3. List in a similar way the objections or disadvantages of such test batteries
4. What is your own attitude with respect to the relative usefulness of such tests of general achievement as compared with a specific survey of a single subject, such as silent reading, as made possible by such a test as the *Iowa Silent Reading Test*?
5. How does the general survey battery fit into the diagnostic and remedial program?
6. Comment on the practical advantages of the arrangement of the items in the *Unit Scales of Attainment*.
7. Who is likely to profit more from the use of general achievement tests, the classroom teacher or the supervisor?
8. Specifically, how may general achievement tests be used to objectify the grade placement and sectioning of pupils in a school system?
9. What is the advantage of the use of the C-score and other standard scores of tests described in this chapter?
10. Secure from your instructor complete sample sets of the *Stanford Achievement Test*, the *Metropolitan Achievement Test*, and the *Unit Scales of Attainment*. After critically examining each test, make a recommendation of one such test battery for use in a general survey in grades 4, 5, and 6. Give the reasons for your choice.

SELECTED REFERENCES

- Allen, Richard D., et al, *Metropolitan Achievement Tests: Supervisor's Manual* Yonkers-on-Hudson, N Y World Book Co., 1935.
- The Cooperative Achievement Tests: A Handbook Describing Their Purpose, Content, and Interpretation* New York Cooperative Test Service, October 1936
- The Cooperative Achievement Tests: Typical Items Illustrating the Form and Content of Representative Tests* New York Cooperative Test Service, October 1938
- Gates, Arthur I., et al, *Modern School Achievement Tests: Manual of Directions* New York Bureau of Publications, Teachers College, Columbia University, 1931
- Gilliland, A R., Jordan, R H., and Freeman, Frank S., *Educational Measurements and the Class-Room Teacher* (Revised Edition), Chapter XVI. New York The Century Co., 1931.
- Iowa Every-Pupil Tests of Basic Skills: Manual of Interpretation*. Boston Houghton Mifflin Co., 1940
- Kelley, Truman L., Ruch, Giles M., and Terman, Lewis M. *Stanford Achievement Tests: Manual for Interpreting*. Yonkers-on-Hudson, N Y World Book Co., 1940
- Nelson, M J., *Tests and Measurements in Elementary Education*, Chapter X New York The Cordon Co., 1939.

Progressive Achievement Tests Manual of Directions. Los Angeles: California Test Bureau, 1937.

Unit Scales of Attainment Manual of Directions and Interpretations. Minneapolis Educational Test Bureau, 1939

Webb, L. W., and Shotwell, Anna Markt, *Testing in the Elementary School*, Chapter XIX. New York: Farrar and Rinehart, Inc., 1939.

CHAPTER XXII

SUMMARIZING THE RESULTS OF TESTING

This chapter gives consideration to the following points in the summarization of test results.

- a.* Statistical procedures needed in summarizing test results.
- b.* Tabulation of test scores.
- c.* Common measures which express typical achievement.
- d.* Common measures of spread or variation.
- e.* Inter-relationships of test scores and abilities.
- f.* Methods of using test scores to place pupils in order of ability.

Introduction. It is common knowledge that a wide range of accomplishment may be expected from the different individuals in a given class. This means that scores representing objective measures of achievement in the classroom will vary widely. Since the human mind is not able to grasp and hold numerous unlike facts in isolation, accurate description of test results depends upon their statistical summarization. Summaries and descriptions of this type need not disturb the student, for after all most of these elementary statistical procedures are simple. The main difficulties are the learning of a new and different type of vocabulary, and the revival of a few relatively simple arithmetical skills.

The use of statistical methods in the analysis of test results is directly in line with good scientific technique. Scientific method in handling test results involves:

- (1) The collection of facts. Within the limits of accuracy of the tests used, the test scores may be said to represent facts.
- (2) The classification and organization of the facts. Simple statistical practices of grouping and tabulating data are utilized for this purpose.
- (3) The further reduction and analysis of the data. Such common statistical procedures as determining measures of central tendency, variability, and relationship are required at this point.

The most important statistical techniques from the standpoint of the frequency of their use in the classroom are abilities to: (1) classify and tabulate data, (2) determine the common measures of central tendency, (3) determine the common measures of spread or variability, (4) determine the relationship between two groups of data, (5) utilize simple graphic methods in the presentation of facts, and (6) secure derived scores and use them in the interpretation of results.

The discussion of these items constitutes the major portions of this and the following chapter. This relatively large amount of emphasis is given to these points for two reasons: (1) Successful and satisfactory work with test results can be expected only when the person using them is adequately equipped to understand and interpret them. Such abilities are dependent upon a reasonable mastery of these elementary statistical techniques. (2) Current educational literature in practically all fields is literally filled with the terms and the techniques discussed here. Reports of progress in education are dependent upon statistical methods. If the teacher and the student are to keep up to date educationally, they must develop the ability to read with understanding the statistical discussions in current educational literature.

. . I. CLASSIFYING AND TABULATING TEST SCORES

Need for a Method of Grouping Data. The very fact that people are unlike physically and mentally gives rise to the need for statistical methods in psychology and education. For example, it may be observed readily from Table XXV that there are great differences in the scores made by the thirty-seven pupils who took a certain arithmetic test. However, it requires rather careful scrutiny to determine that the highest and lowest scores are respectively 72 and 24, while very little further information can be obtained from these scores without rearranging them.

The relatively simple practice of arranging test scores in order of size from highest to lowest or the reverse is helpful, however. Table XXVI reproduces the arithmetic test scores of the thirty-seven pupils in descending order. It now is more easily apparent than from Table XXV that the

TABLE XXV
ARITHMETIC TEST SCORES OF 37 SEVENTH-GRADE PUPILS IN
ALPHABETICAL ORDER OF PUPILS' NAMES

72	45	57	70	66	53	34	32
46	58	24	66	53	36	55	46
63	28	68	52	40	40	48	
30	68	25	38	64	60	55	
71	43	40	60	50	54	30	

highest and lowest scores are respectively 72 and 24, while it can also rather easily be determined that the middle score, or *mid-score*, is 52.

TABLE XXVI
ARITHMETIC TEST SCORES OF 37 SEVENTH-GRADE PUPILS IN
DESCENDING ORDER

72	66	60	54	48	40	34	25
71	66	58	53	46	40	32	24
70	64	57	53	46	40	30	
68	63	55	52	45	38	30	
68	60	55	50	43	36	28	

Table XXVII shows four consistent ways in which these same thirty-seven scores can be classified into a *frequency distribution*. The first illustration, in which the scores retain their individual identities, may be called a *simple frequency distribution*. The other three illustrations, in which the grouping of scores destroys the individual identities of most of them, are called *grouped frequency distributions*. The first distribution furnishes the basis for obtaining detailed information concerning these scores, but such information would be rather costly in the time required to derive it. The fourth illustration furnishes the basis for obtaining quick but quite unsatisfactory information, for the very rough grouping almost entirely sacrifices even the approximate identity of the individual scores. The second and third illustrations, neither of which demands an undue time expenditure in order to obtain accuracy nor sacrifices accuracy for a saving in time and labor, represent something of a "Golden Mean" between the two extreme methods of handling the scores. The second is somewhat preferable to the third for these data.

In general, a method which results in not fewer than ten or twelve and not more than twenty or twenty-five class-intervals or steps is to be preferred in handling test data.

The method of grouping test scores is not dissimilar to that followed by postal clerks in the distribution of letters and other mail in a large railway postal terminal. Mail designated for certain sections of the country or for certain

TABLE XXVII
ARITHMETIC TEST SCORES OF 37 SEVENTH-GRADE PUPILS
IN FREQUENCY DISTRIBUTIONS

Intervals of 1 Unit				Intervals of 3 Units		Intervals of 5 Units	
Scores	<i>f</i>	Scores	<i>f</i>	Scores	<i>f</i>	Scores	<i>f</i>
72	1	55	2	39		68-72	5
71	1	54	1	38	1	63-67	4
70	1	53	2	37		58-62	3
69		52	1	36	1	53-57	6
68	2	51		35		48-52	3
67		50	1	34	1	43-47	4
66	2	49		33		38-42	4
65		48	1	32	1	33-37	2
64	1	47		31		28-32	4
63	1	46	2	30	2	23-27	2
62		45	1	29		Intervals of 15 Units	
61		44		28	1		
60	2	43	1	27		Scores	<i>f</i>
59		42		26		68-82	5
58	1	41		25	1	53-67	13
57	1	40	3	24	1	38-52	11
56						23-37	8
				71-73	2		
				68-70	3		
				65-67	2		
				62-64	2		
				59-61	2		
				56-58	2		
				53-55	5		
				50-52	2		
				47-49	1		
				44-46	3		
				41-43	1		
				38-40	4		
				35-37	1		
				32-34	2		
				29-31	2		
				26-28	1		
				23-25	2		

large centers from which it is re-distributed is thrown into the proper mail pouches. Each pouch may, however, contain mail addressed to a number of postoffices in the same section of the country. For example, in the Chicago postal terminal, mail addressed to postoffices in the Pacific north-west may be consigned to a certain group of pouches, while mail going to the south-east section of the United States will be consigned to other pouches representing that section of the

country. The number of pouches required depends upon the number of pieces of mail to be distributed and also upon the size of the section of the country which can be most efficiently served by a given pouch. Increasing the number of pouches naturally increases the labor involved in sorting the pieces of mail, but at the same time it increases the accuracy of the distribution. Mail in Chicago might be sorted into two classes—eastbound and westbound. This would introduce a large error, since not all sections of the country would be effectively served by this rough classification. The other extreme, of using at this point a separate pouch for the mail addressed to each post office, would be entirely impracticable.

Steps Involved in the Tabulation of Scores. The foregoing illustration will give the reader a clear conception of the purposes and the problems involved in grouping test scores into a frequency table. The steps of procedure presented in the following paragraphs are for the student's guidance in the preparation of grouped frequency distributions of test scores.

Determine the Range of the Scores. Find the highest score and the lowest score in the series. Find the difference between these scores. The difference is called the *range* (*R*) of the distribution. In the case of the scores given in Table XXV, the range is $72 - 24$, or 48. The range is useful in determining the number of class-intervals or steps to use in the frequency table. This is similar to determining how far apart the sections of the country are which must be served by the postal terminal station, or how many mail pouches should be used.

Determine the Size of the Class-Intervals. Use the range to determine whether the scores should be grouped by units of 1, 3, 5, 7, or 15, i.e., to determine the size of the class-intervals. For the range of 48 found for the scores of Table XXV, the correct grouping is by intervals of 3 units each. This is equivalent to determining the number of pouches needed and their arrangement for the most efficient distribution of mail. The mail cannot be sorted accurately until the number of mail pouches and the exact section of the country served by each have been determined.

TABLE XXVIII
RELATION OF RANGE AND SIZE OF CLASS-INTERVAL

For a Range of	Use a Class-Interval of
25 or less	1
26 to 69	3
70 to 125	5
126 to 175	7
176 or more	15

No special rule relative to the number of steps or class-intervals to be used in a frequency table can be stated. However, it is usually unwise to group data into fewer than ten or twelve class-intervals because of the greater *error of grouping* as the number of steps is decreased. Likewise, it is usually undesirable to use more than twenty or twenty-five class-intervals because of the increased labor involved. The main idea of grouping the scores into approximately twelve to twenty or so steps is to organize the scores into a sufficiently small number of groups that they may be thought about effectively and yet not classify them into so few groups that important differences are covered up or significant errors are introduced.

Set up the Frequency Table. Arrange a data sheet consisting of three columns headed "Class-Interval," "Tabulation," and "Frequency." The first and third columns are often headed by the abbreviations "*c.i.*," and "*f.*"

Determine the Limits of the Class-Intervals and Complete the c.i. Column. Under the heading "Class-Interval," write the limits of the intervals into which the scores are to be grouped, beginning at the top with the interval which will include the highest score and continuing downward consistently to include at the bottom the interval which will include the lowest score. To do this, find the multiple of the class-interval which is closest to or equal to the highest score in the series. This number is the mid-point of the highest class-interval. Then establish the integral, i.e., whole-number, limits of the interval equal distances above and be-

low this mid-point, so that the distance between the *integral limits* is one scale unit less than the size of the class-interval. The *real limits* of the intervals are then found by continuing upward .5 of a score unit above the higher integral limit and .5 of a score unit below the lower integral limit. The other intervals are then determined merely by counting downward from the highest interval.

For example, to return to the scores of Table XXV, for which it has been determined above that the grouping should be by class-intervals of 3 units each, the highest score, 72, is exactly divisible by 3, so 72 is the mid-point of the highest class interval. Then 73 and 71 will become the higher and lower integral limits respectively, and 73.5 and 70.5 will become the higher and lower real limits respectively, of the class-interval. The next lower interval will have a mid-point of 69, integral limits of 68 and 70, and real limits of 67.5 and 70.5. The first three columns of Table XXVII show these various points for each interval for the entire distribution based on the scores of Table XXV. In actual work with test scores, however, integral limits only are usually shown in a frequency distribution, except possibly for the upper or lower intervals.

The above directions and illustration will be clarified if the terms are reviewed and defined. The *class-interval* or *step* is the group, or compartment, within the limits of which given scores are assigned. The *mid-point* of the step or interval is a point mid-way between the upper and the lower limits of the step. The *integral limits* are the limits or boundaries of the interval in terms of whole numbers. The *real limits* are the actual boundaries of the interval. For convenience in tabulation it is found desirable to choose the limits of the step in such a way that the mid-point is a whole number. This, of course, makes it necessary that the upper and lower real limits of the step be fractional values whenever odd-sized steps are used. Many statisticians prefer this method because they recognize that, although test scores are usually not given in fractional values, a score of a certain value, say 72, might if the measurement were more accurate equally well represent a score a fraction above 72 or a fraction below 72. A score of 72, then, comes to represent

any score between 71.5 and 72.5. This method has the merit of furnishing a natural location on the scale for all scores expressed in whole numbers.¹

A representation of the mid-point, integral limits, and real limits of the highest class-interval in the distribution of Table XXIX is given in Figure 24, so that the student

TABLE XXIX

ARITHMETIC TEST SCORES OF 37 SEVENTH-GRADE PUPILS
IN A GROUPED FREQUENCY DISTRIBUTION

Class-Interval (<i>c.i.</i>)			Tabulation	Frequency (<i>f</i>)
Integral Limits	Real Limits	Mid-Point		
71-73	70.5-73.5	72	///	2
68-70	67.5-70.5	69	///	3
65-67	64.5-67.5	66	///	2
62-64	61.5-64.5	63	///	2
59-61	58.5-61.5	60	///	2
56-58	55.5-58.5	57	///	2
53-55	52.5-55.5	54	////	5
50-52	49.5-52.5	51	///	2
47-49	46.5-49.5	48	///	1
44-46	43.5-46.5	45	///	3
41-43	40.5-43.5	42	///	1
38-40	37.5-40.5	39	///	4
35-37	34.5-37.5	36	///	1
32-34	31.5-34.5	33	///	2
29-31	28.5-31.5	30	///	2
26-28	25.5-28.5	27	///	1
23-25	22.5-25.5	24	///	2
				<i>N</i> = 37

¹ This represents one of the two most common assumptions made in the statistical work concerning the meaning of a test score. The other widely used statistical method assumes that the true score, of which the test score actually obtained is only an estimate, is not likely to be less than the obtained score but may lie anywhere between the obtained score and a score one unit greater. For example, a score of 72 represents a true score somewhere from 72 to 72.9999.

The authors believe the method used in this volume represents the more sound assumption concerning the meaning of a test score. However, instructors preferring to use the other method can do so easily by shifting each mid-point and each real limit of a class-interval 5 of a score point upward from the values given in this and the following chapter. For example, the real limits of 70.5 and 73.5 for the interval 71-73 would become 71.0 and 73.9999. . . in the other method, and the mid-point of 72 for the same interval would become 72.5 in the other method.

may understand thoroughly the meaning of these terms. It should be clear that the real limits are the points at which adjacent intervals touch, so that the lower real limit of one interval is the higher real limit of the interval next below it.

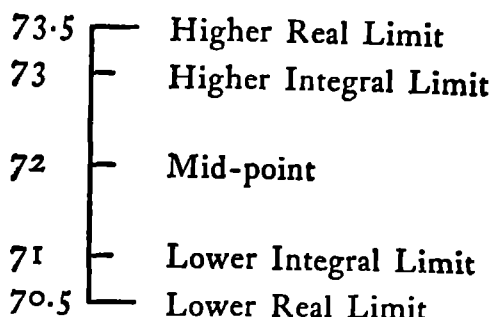


FIGURE 24. MID-POINTS AND LIMITS OF CLASS-INTERVAL

Tabulate the Scores. Begin with the first score in the original list of scores. Determine in which class-interval this score will be included. Place a tally mark in the *Tabulation* column opposite the appropriate class-interval. Make another tally mark for the second score opposite the interval in which it is included. Continue thus until a tally mark has been made for each score in the series. Make each fifth mark in any interval a slanting mark across the preceding four tally marks. Complete the frequency distribution by totaling the tally marks in each row and writing the proper number for that row in the *Frequency* column, and then by obtaining the sum of these frequencies. This sum, N , should equal the total number of original scores.

Summary of Steps in the Tabulation of Scores. The classroom teacher will very often find the construction of a frequency table unnecessary in his experience with tests, since he usually works with small numbers of cases and can check the scores from the papers themselves. However, there are many occasions when the frequency distribution is necessary. It is an effective way of recording and preserving the results of using tests in the classroom. It makes possible a number of short cuts in the calculation of certain statistical measures

useful in interpreting test results. The methods by which these measures are computed are given in succeeding sections of this chapter. This section summarizes in concise form the steps necessary in making a frequency table.

- (1) Determine the range (R) by subtracting the lowest score from the highest score.
- (2) Decide upon a number of score units per class-interval of 1 if the range is 25 or less, of 3 if the range is 26 to 69, of 5 if the range is 70 to 125, of 7 if the range is 126 to 175, and of 15 if the range is 176 or more.
- (3) Set up a frequency table with the headings c , T , and f .
- (4) Determine the limits of the class-intervals so their mid-points are divisible by, and the distance between their real limits are equal to, the number of score units they contain, and write their integral limits in descending order in the c column so that the highest and lowest intervals will provide for the highest and lowest scores.
- (5) Tabulate the scores by placing a tally mark under T in the row which properly indicates the position of each score, carry across the total of the tally marks in each row to the f column, and sum the frequencies in the f column to obtain the number of scores (N).

EXERCISES IN TABULATING TEST SCORES

Problem 1

Tabulating Algebra Test Scores

Determine the range, set up a frequency table by following the instructions given above on this page, and tabulate the algebra quiz scores of 27 Ninth-Grade pupils listed below. Do all of your work on the left half of a sheet of paper and save it for further use.

24, 5, 6, 13, 17, 14, 12, 9, 22, 18, 17, 10, 11, 9, 11, 14, 13, 12, 8, 15, 15, 16, 14, 15, 18, 20, 14

Problem 2

Tabulating History Test Scores

Determine the range, set up a frequency table by following the instructions given on this page, and tabulate the history test scores of 30 Tenth-Grade pupils listed below. Do all of your work on the left half of a sheet of paper and save it for further use.

51, 8, 24, 27, 30, 37, 20, 17, 23, 24, 35, 24, 28, 34, 28, 40, 11, 33, 19, 31, 18, 34, 35, 15, 20, 33, 33, 42, 35, 36.

Problem 3

Tabulating Reading Test Scores

Determine the range, set up a frequency table by following the instructions given on page 502, and tabulate the reading test scores of 50 Eighth-Grade pupils listed below. Do all of your work on the left half of a sheet of paper and save it for further use.

18, 43, 48, 58, 44, 53, 72, 22, 44, 38, 61, 44, 55, 71, 14, 51, 60,
39, 41, 68, 2, 46, 53, 50, 59, 39, 47, 59, 37, 45, 56, 16, 39, 45,
41, 29, 28, 45, 37, 42, 40, 23, 37, 19, 42, 25, 33, 9, 34, 26.

II. MEASURES OF CENTRAL TENDENCY

The grouping of test scores into frequency tables is one step in the process of condensing them so that they can be analyzed and interpreted. However, a further step must be taken before it is possible to describe the data. Some single term or value which is representative of the entire table must be found. Since these values which may be taken to represent an entire distribution of scores are usually found near the center of the data when arranged in order of size, they are commonly called *measures of central tendency*. The three common measures of central tendency are: (1) the arithmetic mean, (2) the median, and (3) the mode. Of these three measures of central tendency, the median and the arithmetic mean are used almost exclusively in educational measurements, and are accordingly the only ones emphasized in this discussion.

Computing the Arithmetic Mean from Ungrouped Data.

The arithmetic mean is the best known and the most widely used measure of central tendency. Indeed, the word "average" is thought by many persons to designate the arithmetic mean, although the arithmetic mean is only one of several "averages."

Practically everyone knows how to find and use the so-called average or arithmetic mean. It is commonly defined as the measure resulting from dividing the sum of the measures in the distribution by the number of measures. Thus the arithmetic mean of the scores 93, 90, 89, 88, and 86 is $446 \div 5 = 89.2$. The value of this measure lies in the fact that it lends itself to describing by means of a single

term a group of widely varying scores or measures. It expresses in very compact form one specific fact about the scores in which each single score has a part. On this account it is one of the basic statistical measures of central tendency.

Computing the Arithmetic Mean (A. M.) from Grouped Data. The arithmetic mean can also be readily found from a frequency distribution. In order to make the procedure somewhat more definite it is advisable to redefine the term *arithmetic mean*. When considered from this point of view *the arithmetic mean is defined as a point on the scale such that the sum of the deviations of the values larger exactly equals the sum of the deviations of the values smaller than it is*. Expressed in physical terms, it may be thought of as the point at which the fulcrum must be placed in order to balance the scale, when it is considered as a beam of varying thickness or density. This point may be determined experimentally or by mathematical calculation. Without regard to the method employed, the fulcrum must be so placed that the moments of the forces on one side are exactly equalled by the moments of the forces on the other side.

Figure 25 illustrates the principle of moments of force by a beam in balance when a weight of one pound is suspended three feet from the fulcrum and a weight of three pounds is suspended one foot from the fulcrum.

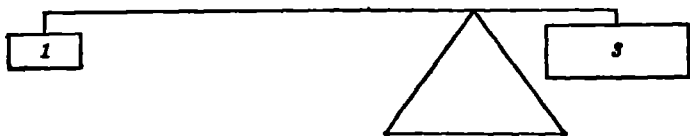


FIGURE 25 ILLUSTRATING THE PRINCIPLE OF MOMENTS OF FORCES

The parallel between the physical lever and the mathematical calculation of the arithmetic mean is quite close. The problem in each case is to balance the forces on either side of a point to be determined. If the physical lever is out of balance, the correction is made by moving the fulcrum in the direction of the heavier end until equilibrium is established. In calculating the arithmetic mean a sort of trial

balance is taken. If the moments of force are too great on one side, the point of rotation is similarly moved in the direction of the heavy end until the difference between these two forces becomes zero.

This may be aptly illustrated by a procedure which classroom teachers have undoubtedly frequently used. It is necessary to obtain the average of five class marks for various periods through the school year in arriving at the final mark. These marks may be taken for illustration as 93, 90, 89, 88, 86. By inspection it may be seen that 89 is approximately the correct mean. The 90 is one point too large, the 93 is four points too large. In a corresponding way 88 is one point too small and 86 is three points too small. The total of the differences above the assumed mean is five and the total of the differences below the assumed mean is four; therefore the assumed mean of 89 is too small by the amount of this difference (1) divided by the number of cases (5). This is equal to .2, so the mean is 89.2. This checks exactly with the mean found by the method of totaling the measures and then dividing by the number of measures, given on page 503.

This method of computing the arithmetic mean will now be applied to the grouped frequency distribution given in Table XXIX for thirty-seven arithmetic test scores.

Assume a Mean. The mid-point of a class-interval near the middle of the frequency distribution should be taken as the assumed mean. This class-interval is usually chosen so that it fairly closely approximates the arithmetic mean. As a matter of fact, however, the results will be the same regardless of the particular interval whose mid-point is chosen as the assumed mean. In the illustration of Table XXX, the guess has been made that the arithmetic mean will fall in or near the interval 50-52, so the assumed mean is 51. For reasons which will be discussed in a later section of this chapter, it is common practice in computing the arithmetic mean to assume that all scores in each step have the value of the mid-point of the step.

Lay off Deviations from the Assumed Mean. Fill in the *d* column by assigning a deviation of 0 to the class-interval in which the assumed mean is located and then count-

TABLE XXX

COMPUTATION OF THE ARITHMETIC MEAN FOR THE GROUPED
FREQUENCY DISTRIBUTION OF 37 ARITHMETIC TEST SCORES

Class-Interval (<i>c</i> i.)		Fre- quency (<i>f</i>)	Devia- tion (<i>d</i>)	<i>fd</i>	1 Assume a mean (51) 2 Lay off deviations from the assumed mean in the <i>d</i> column 3. Fill in the <i>fd</i> column 4 Find the algebraic sum of the values in the <i>fd</i> column $\Sigma fd = +65 + (-83) =$ $65 - 83 = -18$ 5. Divide Σfd by <i>N</i> . $\frac{\Sigma fd}{N} = \frac{-18}{37} = -.49$ 6 Multiply $\frac{\Sigma fd}{N}$ by <i>c</i> : $c = 3 \times -.49 = -1.47$ 7. Algebraically add <i>c</i> to the assumed mean $51 + (-1.47) =$ $51 - 1.47 = 49.53 (AM)$
Integral Limits	Mid- Point				
71-73	72	2	+7	+14	
68-70	69	3	+6	+18	
65-67	66	2	+5	+10	
62-64	63	2	+4	+8	
59-61	60	2	+3	+6	
56-58	57	2	+2	+4	
53-55	54	5	+1	+5	
50-52	51	2	0	0	
47-49	48	1	-1	-1	
44-46	45	3	-2	-6	
41-43	42	1	-3	-3	
38-40	39	4	-4	-16	
35-37	36	1	-5	-5	
32-34	33	2	-6	-12	
29-31	30	2	-7	-14	
26-28	27	1	-8	-8	
23-25	24	2	-9	-18	
		<i>N</i> = 37		(-18)	

ing both upward and downward from that interval by units. Deviations below the assumed mean have negative signs. This is equivalent to showing the number of class-intervals by which each interval deviates from the one containing the assumed mean, and also its direction from that interval, so the deviations are said to be stated in terms of class-intervals.

*Fill in the *fd* Column.* Multiply each frequency by its corresponding deviation and place the results in the *fd* column. Products below the assumed mean will have negative signs.

*Find the Algebraic Sum of the Values in the *fd* Column (Σfd).* Add the *fd* values algebraically by obtaining the

Multiply $\frac{\sum fd}{N}$ by the size of the Class-Interval. Since each step in this table is based on three units, it is necessary to multiply $-.49$ by 3 in order to turn the correction into scale units. The resulting value of -1.47 represents the exact amount by which the assumed mean must be corrected.

Algebraically add $c.t. \times \frac{\sum fd}{N}$ to the Assumed Mean. The arithmetic mean results from the algebraic addition of the assumed mean and the correction (c). As the sign of the correction in the illustration is negative, the correction is subtracted from the assumed mean. Therefore, the arithmetic means is $51 - 1.47$, or 49.53 (*A.M.*). This step is equivalent in the illustration of Figure 26 to moving the fulcrum 1.47 score units downward to bring the beam into balance.

Summary of Steps in Computing the Arithmetic Mean of a Grouped Frequency Distribution. The steps below summarize rather concisely the steps of procedure outlined in detail above for computing the arithmetic mean (*A.M.*) from a grouped frequency distribution.

- (1) Assume a mean as the mid-point of an interval near the middle of the distribution.
- (2) Lay off deviations from the assumed mean in the d column by assigning a value of 0 to the interval in which the assumed mean lies and counting both upward and downward from that interval by units. Deviations below the assumed mean will have negative signs.
- (3) Fill in the fd column by multiplying each frequency by its corresponding deviation. Products below the assumed mean will have negative signs.
- (4) Find the algebraic sum of the values in the fd column by obtaining the sum of the positive values, the sum of the negative values, and then the numerical difference between them. The sign of the difference will be that of the larger value.
- (5) Divide the result of step (4) by N , retaining the proper sign.
- (6) Multiply the result of step (5) by the size of the class-interval, retaining the proper sign, to obtain the correction (c).
- (7) Algebraically add the correction to the assumed mean to obtain the arithmetic mean (*A.M.*). Add the correction to the assumed mean if its sign is positive and subtract it from the assumed mean if its sign is negative.

PROBLEMS IN COMPUTING THE ARITHMETIC MEAN

Problem 4

Computing the Arithmetic Mean

Complete the steps in the computation of the arithmetic mean in the frequency table given below.

Class-Intervals		<i>f</i>	<i>d</i>	<i>fd</i>	Class-interval = Assumed mean = Σfd = $\frac{\Sigma fd}{N}$ or <i>c</i> = Correction $\times 3$ = $A.M. = \text{Assumed mean} + \left(\frac{\Sigma fd}{N} \times 3 \right)$ = = 18.17
Integral Limits	Mid-points				
44-46	45	1			
41-43	42	0			
38-40	39	0			
35-37	36	1			
32-34	33	4			
29-31	30	1			
26-28	27	1			
23-25	24	8			
20-22	21	6			
17-19	18	6	0	0	
14-16	15	6			
11-13	12	11			
8-10	9	5			
5-7	6	3			
2-4	3	0			
0-1	0	1			
		<i>N</i> =			

Problem 5

Computing the Arithmetic Mean

Compute the arithmetic mean for the frequency table prepared for Problem 2, page 502. (*A.M.* = 28.20)

Problem 6

Computing the Arithmetic Mean

Compute the arithmetic mean for the frequency table prepared for Problem 3, page 503. (*A.M.* = 41.10)

Computing a Counting Median or Mid-Measure. Early workers with tests popularized the practice of taking the score of the middle paper of a pile of test papers arranged

in order of size of scores as the expression of the central tendency of the group. The ease with which this so-called median is found has appealed to the classroom teacher. For a long time this was called the *median*. However, more recent workers with tests have recognized that the score of the middle paper in a pile of test papers stacked in order of size of scores is not the same as the middle point on the scale of a frequency table of these same scores. Accordingly a distinction is now made between the score found on the middle paper of a pile of stacked papers and the median proper. The score of the middle paper of a pile of test papers arranged in systematic order is called the *mid-measure* to distinguish it from the *median*, which is the corresponding value when the data are grouped in a frequency distribution. Thus the mid-measure is a counting median found from ungrouped data. The median is computed only from tabulated data. The method of computing the mid-measure is illustrated by referring to the data given in Table XXVI, page 495, where these scores are arranged in descending order by columns. *The mid-measure is the score of the middle paper when the number of cases is odd, or the average of the two scores nearest the middle when the number of cases is even.* In this case the number of papers or scores is 37 (odd). Thus the mid-measure is 52, a score such that there are just as many equal to or larger as there are equal to or smaller than it is.

Computing the Median (Mdn.) from Grouped Data. By definition, the mid-measure and the median are quite similar, the main distinction being that the mid-measure is designated as an actual score on a certain paper (or the average of the scores on the two middle papers) while the median is defined directly in terms of a point on the scale of the frequency table on which it is based. *The median is a point on the scale such that 50 percent of the cases in the distribution are above it and 50 percent of the cases are below it.*

The method of computing the median from a grouped frequency distribution is presented below and illustrated in Table XXXI for the same group of arithmetic test scores used previously in this chapter.

Divide the Number of Cases by 2. The number of cases is

divided by two in order to know how many of the cases fall below the median. For this illustration the half-sum, or $\frac{N}{2}$, is $37 \div 2$, or 18.5.

Count Upward into the Distribution to Obtain the Sub-Total. Count upward into the distribution, adding the frequency for each interval, until exactly the half-sum or a number as closely approaching it as possible without exceeding it is reached. Thus, in the illustration, $2 + 1 + 2 + 2 + 1 + 4 + 1 + 3 + 1 = 17$. If the two scores in the interval 50-52 were added, the result, 19, would exceed the half-sum. The median, therefore, lies somewhere in the interval 50-52, for less than half of the scores lie below that interval and less than half of the scores lie above it.

TABLE XXXI

COMPUTATION OF THE MEDIAN FOR THE GROUPED FREQUENCY
DISTRIBUTION OF 37 ARITHMETIC TEST SCORES

Class-Interval (<i>c</i> 1)		Fre- quency (<i>f</i>)	1 Divide the number of cases by 2. $\frac{N}{2} = \frac{37}{2} = 18.5.$
Integral Limits	Real Limits		
71-73	70.5-73.5	2	2 Count upward into the distri- bution to obtain the sub-total. Sub-total = $2 + 1 + 2 + 2 + 1 + 4 + 1 + 3 + 1 = 17.$
68-70	67.5-70.5	3	
65-67	64.5-67.5	2	
62-64	61.5-64.5	2	
59-61	58.5-61.5	2	3 Determine the correction (Measures) $18.5 - 17 = 1.5.$
56-58	55.5-58.5	2	
53-55	52.5-55.5	5	4. Determine the correction (Proportion) $1.5 - 2 = .75.$
50-52	49.5-52.5	2	
47-49	46.5-49.5	1	5. Determine the correction (Scale Distance) $.75 \times 3 = 2.25.$
44-46	43.5-46.5	3	
41-43	40.5-43.5	1	6. Obtain the median. $49.5 + 2.25 = 51.75. (Mdn.)$
38-40	37.5-40.5	4	
35-37	34.5-37.5	1	
32-34	31.5-34.5	2	
29-31	28.5-31.5	2	
26-28	25.5-28.5	1	
23-25	22.5-25.5	2	
		$N = 37$	

Determine the Correction in Terms of Measures. Subtract the sub-total from the half-sum. This subtraction will give the number of cases in the interval in which the median lies which must be added to the sub-total to obtain the half-sum, and consequently shows how much further counting must continue upward to obtain the median. In the distribution of Table XXXI, this step becomes $18.5 - 17 = 1.5$.

Determine the Correction in Terms of Proportion of the Interval. Divide the result of the preceding step (half-sum—sub-total) by the number of cases in the interval in which the median falls. This will give the proportion of the interval which must be added to lower intervals in order to reach the point below which half of the cases fall. For the illustration of Table XXXI, $1.5 \div 2 = .75$. This step is based on the assumption that the scores in an interval are uniformly distributed in the interval. More will be said about this assumption in a following section.

Determine the Correction in Terms of Scale Distance. Multiply the result of the preceding step by the size of the class-interval so that the correction will be stated as a scale distance. Thus, $.75 \times 3 = 2.25$ for the accompanying illustration.

Determine the Median (Mdn.). Now add the correction in terms of scale distance to the lower real limit of the interval in which the median lies to obtain the median. The correction of 2.25 added to 49.5, or the lower real limit of the interval in Table XXXI which contains the median, gives 51.75 (Mdn.).

Obviously, if the calculations of these steps were made by adding the frequencies down from the top of the distribution the median would be the same. In that case, 18 scores falling above the interval 50-52, the correction of .75 (.5 of a measure, .25 of an interval, .75 in terms of scale units) would be subtracted from 52.5, the top of the step, to give the same result of 51.75 for the median.

Summary of Steps in Computing the Median of a Grouped Frequency Distribution. The steps listed below provide in form for easy use the procedures necessary for computing the median (Mdn.) for a grouped frequency distribution.

- (1) Divide the number of cases by 2 to obtain the half-sum.
- (2) Count upward into the distribution to obtain the sub-total by adding frequencies of intervals until exactly the half-sum or a number as closely approaching it as possible without exceeding it is reached ³
- (3) Determine the correction in terms of measures by subtracting the sub-total from the half-sum.
- (4) Determine the correction in terms of proportion of the interval by dividing the result of step (3) by the number of cases in the interval in which the median falls.
- (5) Determine the correction in terms of scale distance by multiplying the result of step (4) by the size of the class-interval.
- (6) Obtain the median by adding the correction of step (5) to the lower real limit of the interval in which the median falls.

PROBLEMS IN COMPUTING THE MID-MEASURE AND MEDIAN

Problem 7

Finding the Mid-measure

Find the mid-measure or counting median for the data of Problem 1, page 502. (*Mid-measure* = 14)

Problem 8

Finding the Mid-measure

Find the mid-measure for the data of Problem 2, page 502. (*Mid-measure* = 29)

Problem 9

Computing the Median

Compute the median from the frequency table prepared for Problem 2, page 502. (*Medn.* = 28.50)

Problem 10

Computing the Median

Compute the median from the frequency table prepared for Problem 3, page 503. (*Medn.* = 41.94)

Basic Assumptions in Computing Measures of Central Tendency. As has been indicated briefly in a preceding section of this chapter, the assumption concerning the distribution of scores within each class interval varies according to

³ If exactly the half-sum is reached, the median is usually the upper real limit of the interval whose frequency was last added in the counting process. However, if the next higher interval should happen to have a zero frequency, the median is the mid-point of that interval.

which of the measures of central tendency is being computed. Figure 27 shows in parallel graphic representations of the distribution of scores assumed in the computation of the arithmetic mean and the median for several class-intervals near the center of the distribution used for illustrative purposes in the preceding pages.

c. i.	f	Arithmetic Mean		Median	
		Real Limits of Steps	Mid-points of Steps	Real Limits of Steps	Divisions between Measures
56-58	2	58.5 _____	57.0	58.5 _____	58.5
		_____		_____ 2 _____	57.0
53-55	5	55.5 _____	54.0	55.5 _____	55.5
		_____		_____ 1 _____	54.9
		_____		_____ 1 _____	54.3
		_____		_____ 1 _____	53.7
50-52	2	52.5 _____	51.0	52.5 _____	53.1
		_____		_____ 2 _____	52.5
47-49	1	49.5 _____	48.0	49.5 _____	51.0
		_____		_____ 1 _____	49.5
44-46	3	46.5 _____	45.0	46.5 _____	46.5
		_____		_____ 3 _____	45.5
		_____		_____ 2 _____	44.5
		43.5 _____		43.5 _____	43.5

FIGURE 27 ILLUSTRATING THE ASSUMPTIONS CONCERNING THE DISTRIBUTION OF SCORES IN CLASS-INTERVALS IN THE COMPUTATION OF THE ARITHMETIC MEAN AND MEDIAN

It is assumed in the computation of the arithmetic mean that each score in a grouped frequency distribution has the value of the mid-point of the interval in which it is tabulated. This is illustrated by the heavily ruled lines in the left-hand portion of Figure 27. On the other hand, it is assumed in the computation of the median that each score in a grouped frequency distribution expands or contracts in such manner that it shares the scale distance through a class-interval equally with the other measures in the same class-interval. This assumption is illustrated in the right-hand portion of Figure 27. Thus, each of the three measures in the interval 44-46 is assumed to have the value of 45 in computing the arithmetic mean and to occupy one-third of the scale distance through that interval ($1/3 \times 3 = 1$) in computing the median. Again, the five scores in the step 53-55 are assumed in computing the arithmetic mean to be concentrated at 54, the mid-point of the interval, while in computing the median

each of the five scores is assumed to occupy one-fifth of the scale distance through that interval ($1/5 \times 3 = .6$).

This leads to one further important distinction between the arithmetic mean and the median. The mean is algebraic in nature (although the various operations can be stated either in algebraic or in arithmetic terms), while the median is arithmetic in nature. As will be seen later in this chapter, each of these measures is accompanied by one or more measures of variability or dispersion having comparable algebraic or arithmetic origins.

III. MEASURES OF VARIABILITY

Need for Measures of Variability. The measures of central tendency represented by the arithmetic mean and the median are valuable statistical measures but they describe only one characteristic of the data, the tendency of the scores to pile up at or near the middle of the distribution. Descriptions of test results based wholly on one or the other of these measures are incomplete.

TABLE XXXII
DATA SHOWING IDENTICAL MEANS BUT
UNLIKE VARIABILITY

Class A		Class B
122		98
116		96
108		95
101		93
96		90
92		89
89		87
86	A.M.	86
83		85
80		83
76		82
71		79
64		77
56		76
50		74

The two groups of scores presented as Class A and Class B in Table XXXII illustrate this situation very clearly. The means of the two series of scores are identical, each being 86. The *range* of the scores for Class A is 72 ($122 - 50$), which is exactly three times the range ($98 - 74 = 24$) of scores for Class B. Even the most inexperienced teacher or student must recognize that very different ranges of ability are present in these two classes and that correspondingly different instructional problems are presented to the teacher.

A graphic illustration based on other data showing the unlikenesses which may appear in distributions having the same mean is given in Figure 28.

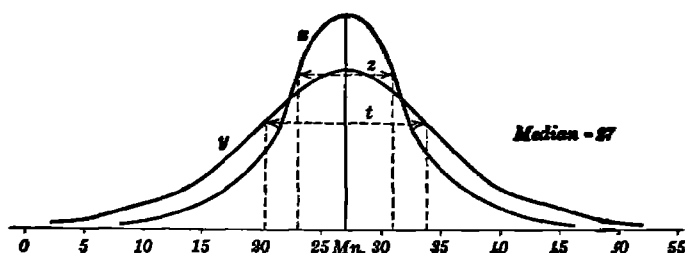


FIGURE 28 ILLUSTRATING THE NEED FOR MEASURES OF VARIABILITY

It is a common practice to show frequency distributions in graphic form by representing the frequencies at a given point on the **scale in terms of a line erected perpendicular to the base line or scale**. If the tops of a large number of these perpendiculars are connected, the result is a curved line which usually is close to the base line at the ends of the scale but rises quite rapidly from the base line near the middle of the scale. In Figure 28 the curve marked *x* represents the distribution of scores made on a certain test by a class. The closeness of the curve to the base line at the ends or extremes shows that there are relatively few very low and very high scores. The high point of the curve near the middle indicates that a great many pupils made scores near the average. This is typical of situations usually found where considerable numbers of cases are involved.

It will be noted in Figure 28 that, while the middle por-

tion of the curve x rises much higher than the similar portion of the curve y , the extremes of curve x do not go out on the base line in either direction so far as is true of curve y . This flatness or peakedness of the curve is the graphic indication of the variability of the data it represents. The less peaked the curve the greater the variability, other things being equal. It is thus apparent from this illustration that while the means of these two distributions are identical, very greatly different teaching problems are represented. Curve x represents a relatively homogeneous group, while curve y represents a more widely scattered group.

Range as an Expression of Variability. The *range* of scores, that is, the scale distance between the lowest and highest scores in a distribution, is one way of expressing variability. However, it is one of the least reliable measures of variability or dispersion, since it is apparent that it is affected by the fluctuation of the extreme scores.

Quartile Deviation as a Measure of Variability. A much more reliable expression of variation is found in the *semi-interquartile range*, which is really the range of the middle 50 percent of the cases. Thus the operation of the unusual deviation of "sport" cases at the extremes of arrays is eliminated. In actual practice *the range of the middle half of the cases is expressed in terms of one-half of the interquartile range and is called the quartile deviation or Q* . The values of the upper quartile (Q_3) and of the lower quartile (Q_1) correspond respectively to the seventy-fifth percentile and the twenty-fifth percentile. The seventy-fifth percentile is the point on the scale below which 75 percent of the cases lie and the twenty-fifth percentile is the point below which 25 percent of the cases lie. The value Q and the median (Q_2) are distinct, however. One is a measure of variation or spread while the other is a measure of central tendency. The difference between the median (*a measure of central tendency*) and the quartile deviation (Q , *a measure of variability*) is made clear by Figure 28, page 516. The median line is the vertical line labeled *A.M.* which cuts the base line at the value 27 on the scale. The median is a *point* on the scale. The value Q is a *distance* on the scale, and in practice is represented by one-half the length of the lines t

or z in the diagram. The difference in the variability of the two distributions represented by the curves x and y is shown by the relative lengths of the lines t and z . The value Q for the curve x is stated as one-half of the line z , and for curve y it is one-half of the line t .

It has been pointed out that the central tendencies of the groups represented by the curves x and y in Figure 28 are identical, while the spread of ability is much greater in distribution y than in x . Line t , which represents the entire range of the middle 50 percent of the cases in distribution y , or two times the Q for this distribution, takes in a much wider range of scores than does line z , which represents the entire range of the middle 50 percent of the scores in distribution x . All measures of variability indicate in one way or another the average or typical amount of deviation from the average of the group as a whole. Central tendencies reveal the most likely point of balance between the high and the low scores in a distribution. Measures of variability show how the group scatters from this central tendency. A distribution with a large measure of variability scatters more widely from the central tendency than does a distribution with a smaller measure of variability.

A comprehension of this general concept underlies an understanding of practically all of the measures of variability as well as a great many of the statistical techniques yet to be considered in connection with the interpretation of tests.

The quartile deviation (Q) and the standard deviation, or sigma (σ) are representative of the more important measures of variability, and constitute the ones treated intensively in this book.

Computing the Quartile Deviation (Q) from Grouped Data. The method of computing the quartile deviation (Q) is demonstrated by the use of the data given in Table XXV, page 495.

The first quartile, or twenty-fifth percentile, and the third quartile, or seventy-fifth percentile, of a distribution are found in the same manner as is employed in finding the median except that, of course, the numbers of cases counted off above or below the points on the scale are different. For the first quartile, one-fourth or 25 percent of the cases are

counted off below the point taken as Q_1 . For the third or upper quartile, 75 per cent of the cases are counted off below the point taken as Q_3 . This point (Q_3) may be more easily obtained by counting off 25 per cent of the cases from the top of the distribution.

The steps involved in the computation of the quartile deviation really involve only one simple procedure beyond those used in computing the median. The following brief presentation is based on the frequency distribution of Table XXXIII, in which the arithmetical operations are shown.

Determine the Upper Quartile (Q_3). The upper quartile, or 75th percentile, is the point below which 75 percent of the cases lie. Consequently 25 per cent of the cases lie above Q_3 . Therefore, the distance which must be counted down into the distribution is $37 \div 4$, or 9.25 cases. There are 9 cases down to the top of the interval 59-61, so an additional .25 case must be counted downward. This is .13 of a class-interval ($.25 \div 2$), or, in score units, .39 ($.13 \times 3$). Therefore Q_3 is obtained by subtracting .39 from 61.5, the upper real limit of the interval in which Q_3 falls, to obtain 61.11 as the desired value.

Determine the Lower Quartile (Q_1). Eight cases fall below the interval 38-40, so counting upward must continue for 1.25 additional cases, or .31 of an interval ($1.25 \div 4$). In score distance, this is .93 ($.31 \times 3$). When this is added to the lower real limit of the interval in which Q_1 falls, the desired measure is found to have the value of 38.43.

Determine the Quartile Deviation (Q). The distance between Q_3 and Q_1 is the interquartile range. Half of that distance is the *semi-interquartile range*, which is merely another name for the quartile deviation. Therefore $Q = \frac{Q_3 - Q_1}{2}$. For the illustration, this becomes $\frac{61.11 - 38.43}{2} = \frac{22.68}{2} = 11.34$. (Q).

Summary of Steps in Computing the Quartile Deviation of a Grouped Frequency Distribution. The three steps

given below summarize the procedures involved in the computation of Q .

- (1) Determine the upper quartile (Q_3) or 75th percentile.
- (2) Determine the lower quartile (Q_1) or 25th percentile.
- (3) Determine the quartile deviation (Q) by substituting the values obtained in steps (1) and (2) in the following formula and carrying through the simple arithmetic operations $Q = \frac{Q_3 - Q_1}{2}$.

TABLE XXXIII

COMPUTATION OF THE QUARTILE DEVIATION FOR THE GROUPED
FREQUENCY DISTRIBUTION OF 37 ARITHMETIC TEST SCORES

Class-Interval (c.i.)		Frequency (f)	
Integral Limits	Real Limits		
71-73	70.5-73.5	2	1. $\frac{N}{4} = \frac{37}{4} = 9.25$.
68-70	67.5-70.5	3	2. $2+3+2+2=9$.
65-67	64.5-67.5	2	3. $9.25-9=.25$.
62-64	61.5-64.5	2	4. $.25 \div 2 = .13$.
59-61	58.5-61.5	2	5. $.13 \times 3 = .39$.
56-58	55.5-58.5	2	6. $61.50 - .39 = 61.11 (Q_3)$
53-55	52.5-55.5	5	1. $\frac{N}{4} = \frac{37}{4} = 9.25$.
50-52	49.5-52.5	2	2. $2+1+2+2+1=8$.
47-49	46.5-49.5	1	3. $9.25-8=1.25$.
44-46	43.5-46.5	3	4. $1.25 \div 4 = .31$.
41-43	40.5-43.5	1	5. $.31 \times 3 = .93$.
38-40	37.5-40.5	4	6. $37.5 + .93 = 38.43 (Q_1)$
35-37	34.5-37.5	1	
32-34	31.5-34.5	2	
29-31	28.5-31.5	2	
26-28	25.5-28.5	1	
23-25	22.5-25.5	2	
		$N=37$	$Q = \frac{Q_3 - Q_1}{2} = \frac{61.11 - 38.43}{2}$ $= \frac{22.68}{2} = 11.84$.

PROBLEMS IN COMPUTING THE QUARTILE DEVIATION

Problem 11

Computing the Quartile Deviation

Compute the quartile deviation from the frequency distribution prepared for Problem 2, page 502. ($Q=6.53$)

Problem 12

Computing the Quartile Deviation

Compute the quartile deviation from the frequency distribution prepared for Problem 3, page 503. ($Q = 9.34$)

Standard Deviation as a Measure of Variability. Such simple methods of expressing the variation of a distribution of test scores as the range and the quartile deviation (Q) are sufficient for most ordinary statistical situations, but where careful work of an analytical or research type is being done a more exact means of expressing variability must be used. The standard deviation, also called sigma (σ), has many characteristics which make it a useful measure of variability. *The standard deviation is a sort of arithmetic mean of the deviations from the mean of the distribution.* It is a special type of mean of the deviations because of the method used in computing it. In calculating the standard deviation (σ), each individual deviation from the mean is squared, the sum of these squared values is then divided by the number of such deviations, and the square root of the result is then obtained. Restated, *the standard deviation (sigma) is the square root of the mean of the squares of the deviations from the mean of a distribution.* Expressed in symbols the standard deviation becomes $s\sqrt{\frac{\sum fd^2}{N} - c^2}$ where $\sum fd^2$ equals the deviations in the form of the sum of the products of the frequencies at each step by the deviation of each step from the assumed mean, N equals the number of cases in the distribution, c equals the correction as found in calculating the mean, and s equals the size of the class-interval of the distribution in units.

The relation of the standard deviation to the quartile deviation of a distribution is shown in Figure 29. By definition the *quartile deviation* (Q) takes into account the middle 50 percent of the cases. That is, the ordinates (lines erected perpendicular to the base line of the curve) erected at a scale distance equal to Q on either side of the mean or median include 50 percent of the area of the surface between the curve and the base line. In the diagram the lines P and R represent the ordinates erected at a distance equal to Q on

either side of the mean. The lines S and T represent ordinates erected at a distance equal to σ on either side of the mean. The standard deviation (σ) takes into account approximately 68 percent (in a normal distribution 68.26%) of the area of such a distribution. That is, ordinates erected at a distance equal to sigma on either side of the arithmetic mean include approximately two-thirds of the cases in the distribution.

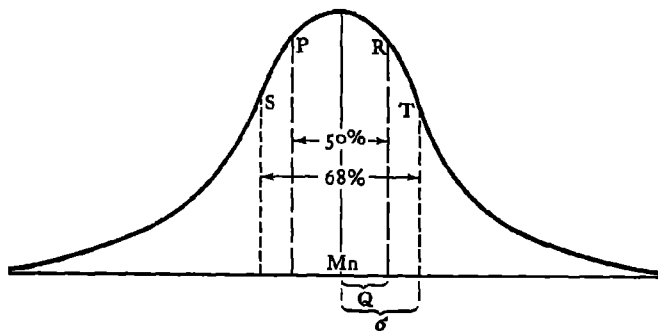


FIGURE 29 COMPARISON OF STANDARD DEVIATION (σ) AND QUARTILE DEVIATION (Q)

The standard deviation (σ) may be computed about any measure of central tendency but it is common practice to consider it as computed about the arithmetic mean only. Unless otherwise indicated, sigma (σ) may be considered as having been calculated about the arithmetic mean.

A further interesting characteristic of the standard deviation is indicated in Figure 29. Mathematically the value *sigma* bears a definite relationship to the curve of the distribution itself. It will be noted that, where any large number of cases or scores are found in a distribution, there is a tendency for the larger portion of the cases to pile up in or near the middle of the distribution. When a normal distribution is presented in graphic form, the result is a symmetrical bell-shaped curve with many cases in the middle and few cases at the extremes. Certain types of these characteristic bell-shaped distributions have come to be called *normal curves*. For these normal curves, formulae have been derived from which such typical curves may be computed if

certain basic data concerning the curve are given. In these formulae *sigma* is one of the values which must be given in order to construct such a curve. Sigma, in the typical formula, represents the distance from the mean at which the curve changes from convex to concave. In Figure 29 the points where the curve changes its character are indicated by the ordinates lettered S and T.

Thus, because of this direct mathematical relationship which the standard deviation bears to the curve of the distribution itself, and the reliable expression of variability which it provides since every deviation in the distribution is considered, the standard deviation is one of the most useful of the measures of variability.

Computing the Standard Deviation from Ungrouped Data. In the computation of the standard deviation from ungrouped data, as in the accompanying illustration, the mean of the distribution must be found. When the data are grouped in a frequency table, it is not strictly necessary for the arithmetic mean to be computed, although it is necessary to go through all but the last step of the process.

The steps in the computation of the standard deviation from ungrouped data are given in detail in connection with data from Table XXXII. The scores used are the same as those which appear for Class A. The mean of the scores for Class A is 86. Thus, a score of 89 deviates from this mean by 3 points. A score of 96 deviates 10 points, etc. (See column *d* in the table.) The standard deviation (σ) is the square root of the mean of the squares of these deviations from the mean of the array of scores. It is necessary, therefore, to square each of these deviations. These are given under the column headed d^2 . Since each deviation appears only once and the data are ungrouped, the formula may be simplified to read $\sigma = \sqrt{\frac{\sum d^2}{N}}$. The sum of the deviations squared ($\sum d^2$) in this case is 6100. The mean of these deviations is therefore 406.67. This value is the mean of the deviations squared. Therefore, to turn it into units of the scale, the square root of this quantity must be taken. This value is 20.17, which is the standard deviation (σ) of this series of scores. The mean of this distribution is 86.

The σ is 20.17. This means that between scores 20.17 points larger and 20.17 points smaller than this mean, approximately two-thirds (68.26%) of the scores will be found.

TABLE XXXIV

COMPUTATION OF THE STANDARD DEVIATION FROM UNGROUPED DATA
(Data for Class A from Table XXXII)

Test Scores	Deviations (d)	Deviations Squared (d^2)	Computations
122	+36	1296	$\sigma = \sqrt{\frac{\sum d^2}{N}}$ $= \sqrt{\frac{6100}{15}}$ $= \sqrt{406.67}$ $= 20.17$ $A.M. = \frac{1290}{15}$ $= 86$
116	+30	900	
108	+22	484	
101	+15	225	
96	+10	100	
92	+6	36	
89	+3	9	
86	0	0	
83	-3	9	
80	-6	36	
76	-10	100	
71	-15	225	
64	-22	484	
56	-30	900	
50	-36	1296	
1290		$\sum d^2 = 6100$	

Computing of the Standard Deviation (S.D. or σ) from Grouped Data. The method of computing the standard deviation from ungrouped data illustrated in Table XXXV may be applied with very few changes to the calculation of sigma from grouped data. A slight change in the general formula is required, for when the scores are grouped in class intervals the deviations of the scores must be considered by groups having the mid-points of the steps in which they are found. This permits the expression of the deviations in steps rather than in units of the scale. The formula for use in calculating the standard deviation when the data are grouped in a frequency distribution is $s = \sqrt{\frac{\sum f d^2}{N} - c^2}$. The

steps in the application of this formula in the calculation of the standard deviation of the scores originally presented in Table XXV will make clear all of the processes involved in finding the sigma of a frequency distribution. The computations themselves are shown in Table XXXV.

The first five steps of procedure for computing the standard deviation are identical with those given above for determining the arithmetic mean.

Assume a Mean. Assume a mean as near as possible to the arithmetic mean of the distribution in order that the correction (c) may be as small as possible. In Table XXXV, as for the computation of the A.M. for the same distribution, the assumed mean is taken as the mid-point of the interval 50-52.

Lay Off Deviations from the Assumed Mean. Lay off deviations above and below the step in which the mean is assumed to lie. Signs of deviations below the assumed mean should be negative.

Fill in the fd Column. Multiply the frequency in each interval by its corresponding deviation and carry the results to the fd column. Take account of signs.

Find the Algebraic Sum of the Values in the fd Column (Σfd). Obtain the sum of the positive fd and the negative fd values separately and then determine the algebraic sum for the entire column. Take proper account of signs of these quantities. In the illustration of Table XXXV, the Σfd is — 18.

Divide the Algebraic Sum of the fd Column by N ($\frac{\Sigma fd}{N}$, or c). Perform this division, taking account of signs, to obtain the correction (c).⁴ For the distribution of Table XXXV, c is —.486. In this computation of the arithmetic mean, this correction was converted into scale units in the following step, but here the correction is left in terms of class-intervals.

⁴ This value, prior to the application of the correction, c^2 , is frequently designated as the *root-mean-square-deviation* or S^2 . The correction, c (squared to make it comparable to the S^2) is always deducted from the root-mean-square-deviation because this value, S^2 , is always too large. If c has any value more than zero, it means that the assumed mean and the arithmetic mean do not coincide, therefore, any deviations computed about it are too large and the correction c must be subtracted to compensate for this difference.

TABLE XXXV

COMPUTATION OF THE STANDARD DEVIATION FOR THE GROUPED
FREQUENCY DISTRIBUTION OF 37 ARITHMETIC TEST SCORES

Class-Interval (c)		Fre- quency (f)	Devia- tion (d)	fd	fd^2	1. Assume a mean (51) 2 Lay off deviations in the d column 3. Fill in the fd col- umn 4 Find Σfd (-18) 5. Find $\frac{\Sigma fd}{N}$ (-.486) 6 Square the correc- tion (.236) 7 Fill in the fd^2 col- umn. 8 Find Σfd^2 (826) 9. Find $\frac{\Sigma fd^2}{N}$ (22.324) 10 Subtract c^2 from $\frac{\Sigma fd^2}{N}$ (22.088) 11. Obtain $\sqrt{\frac{\Sigma fd^2}{N} - c^2}$ (4.70) 12 Obtain $s = \sqrt{\frac{\Sigma fd^2}{N} - c^2}$ (14.10) S D
Integral Limits	Mid- Point					
71-73	72	2	+7	+14	98	
68-70	69	3	+6	+18	108	
65-67	66	2	+5	+10	50	
62-64	63	2	+4	+8	32	
59-61	60	2	+3	+6	18	
56-58	57	2	+2	+4	8	
53-55	54	5	+1	+5	5	
50-52	51	2	0			
47-49	48	1	-1	-1	1	
41-46	45	3	-2	-6	12	
41-43	42	1	-3	-3	9	
38-40	39	4	-4	-16	64	
35-37	36	1	-5	-5	25	
32-34	33	2	-6	-12	72	
29-31	30	2	-7	-14	98	
26-28	27	1	-8	-8	64	
23-25	24	2	-9	-18	162	
		$N=37$		(-18)	(826)	

Square the Correction (c^2). Square the correction, in conformance with the formula given above for the standard deviation.⁵ This value becomes .236 for the distribution of Table XXXV. The c^2 will also be a positive value.

Fill in the fd^2 Column. Multiply each value in the fd column by its corresponding deviation (d), and place the results in the fd^2 column. All signs will be positive.

Obtain the Sum of the fd^2 Column. Obtain the sum of the fd^2 values. The sign will, of course, always be positive. Σfd^2 for Table XXXV is 826.

Divide the Sum of the fd^2 Column by N . $\left(\frac{\Sigma fd^2}{N}\right)$. Di-

⁵ The student will find the tables of squares and square roots such as are given on pages 476-85 of Garrett's *Statistics in Psychology and Education* (Second Edition) or Ours' *Statistical Methods in Educational Measurements*, pages 299-305, very helpful in his work from this point on. Ability to use a slide rule or a mechanical calculator will also speed up the work as well as make it more accurate.

vide the Σfd^2 by N . The result is 22.324 for the distribution of Table XXXV.

Subtract c^2 from $\frac{\Sigma fd^2}{N} \cdot \left(\frac{\Sigma fd^2}{N} - c^2 \right)$. Subtract the square of the correction from $\frac{\Sigma fd^2}{N}$, to account for the deviation of the assumed mean from the arithmetic mean. The result for the accompanying illustration is $22.324 - .236 = 22.088$.

Obtain the Square Root of $\frac{\Sigma fd^2}{N} - c^2 \cdot \left(\sqrt{\frac{\Sigma fd^2}{N} - c^2} \right)$.

Obtain the square root of the value resulting from the above step of procedure to obtain the standard deviation in terms of class-intervals. The square root of 22.088 to two decimal places is 4.70.

Multiply $\sqrt{\frac{\Sigma fd^2}{N} - c^2}$ by the size of the Class-Interval.

$\left(s \sqrt{\frac{\Sigma fd^2}{N} - c^2} \right)$. Multiply the standard deviation in units of class-intervals by the size of the step (s) to put the standard deviation into score units. For the illustration of Table XXXV, the result is 14.10 (*S.D.* or σ).

This means that, if ordinates are erected at a distance of 14.70 units on either side of the arithmetic mean of this distribution, approximately 68 percent of the distribution will be found between these ordinates. Of course this will not be strictly true, since no small frequency table such as the one used here will be sufficiently symmetrical to permit the cases to fall in exactly that fashion. Roughly the *S.D.* expresses the range of the middle 68 percent of the cases.

The arithmetic mean of the data used in this illustration was shown above to be 49.53. The standard deviation (σ) in points on the scale is 14.70. Ordinates erected 14.70 points above and below this arithmetic mean fall at 64.23 and 34.83 on the test scale. Within these points will be found approximately 68 percent of the measures in the distribution.

Summary of Steps in Computing the Standard Deviation of a Grouped Frequency Distribution. The steps of procedure to be used in computing the *S.D.* of a grouped frequency distribution are as follows:

- (1) Assume a mean as the mid-point of an interval near the middle of the distribution
- (2) Lay off deviations from the assumed mean in the d column by assigning a value of 0 to the interval in which the assumed mean lies and counting both upward and downward from that interval by units. Deviations below the assumed mean will have negative signs.
- (3) Fill in the fd column by multiplying each frequency by its corresponding deviation. Products below the assumed mean will have negative signs
- (4) Find the algebraic sum of the values in the fd column by obtaining the sum of the positive values, the sum of the negative values, and then the numerical difference between them. The sign of the difference will be that of the larger value.
- (5) Divide the result of step (4) by N , retaining the proper sign
- (6) Square the correction (c^2). The result will always be positive.
- (7) Fill in the fd^2 column. Multiply each value in the fd column by its corresponding deviation (d), and place the results in the fd^2 column. All signs will be positive.
- (8) Obtain the sum of the fd^2 column. The sign will, of course, always be positive.
- (9) Divide the sum of the fd^2 column by N . $\left(\frac{\sum fd^2}{N}\right)$.
- (10) Subtract c^2 from $\frac{\sum fd^2}{N}$. $\left(\frac{\sum fd^2}{N} - c^2\right)$.
- (11) Obtain the square root of the result of step (10).
- (12) Multiply the result of step (11) by the size of the class-interval to obtain the standard deviation ($S.D.$ or σ).

PROBLEMS IN COMPUTING THE STANDARD DEVIATION

Problem 13

Computing the Standard Deviation

Compute the standard deviation from the frequency distribution prepared for Problem 2, page 502. ($S.D. = 9.21$)

Problem 14

Computing the Standard Deviation

Compute the standard deviation from the frequency distribution prepared for Problem 3, page 503. ($S.D. = 15.50$)

IV. THE RELATIONSHIP OF TEST SCORES

Need for Measures of Relationship. The discussion of this chapter up to this point has concerned itself with two

types of descriptions of test scores: (1) measures of central tendency, and (2) measures of variability. In connection with the more critical examination and interpretation of test results, it often becomes necessary for the teacher or the student to go somewhat beyond the description afforded by these simple measures. Often he must select for testing purposes the one test of a series which most nearly measures the desired ability. The method followed in such a case involves finding the relationship between the several tests under consideration and the ability to be measured. This is called the method of *correlation*. It is the method applied when two or more measures of the same individuals are being studied with the view to finding the degree of relationship between the two or more sets of measures. Sometimes the method of correlation is mistakenly used to attempt to discover causes operating to produce certain effects. There is nothing in the method or the result of computing a correlation which indicates definitely which of the factors is a cause and which is an effect, or whether both of the factors may be affected by other variables.

The Coefficient of Correlation. In the expression of relationships, as in the other statistical measures, it is desirable that this relation between two series of variables be expressed in a single objective or mathematical value. A number of different ways have been proposed for the derivation of these expressions of relationship, but no one of these methods has produced a term which is both objective and easily interpreted. Methods have been proposed for computing relationships in terms of the correspondence between rank positions of scores, and in terms of the percentage of the scores falling within a specified unit of variability of each other. In general these procedures lack sufficient exactness to warrant their common use in the analysis of test results. The student who is interested in these different methods of revealing relationships will find them discussed in certain of the treatments on statistical methods listed in the references at the end of this chapter.

It is probably useless for the student of educational measurements to attempt to master more than one method of computing the coefficient of correlation. The one method

considered here, the *Pearson Product-Moment Method*, is by far the most common and is, on the whole, the basic method used in educational investigations. This method, while somewhat complicated and difficult because of the large number of different calculations to be made, really involves very little that the student has not already mastered.

The Pearson Product-Moment coefficient of correlation, indicated by r , is a single numerical index which expresses the extent to which the pairs of corresponding measures of two variables tend to deviate similarly from their respective arithmetic means. The values of r may vary all the way from $+1.0$, indicating perfect positive relationship, through all of the possible decimals to zero (0), indicating no relationship whatever, to -1.0 , indicating a perfect negative relationship. The following are illustrations of a positive relationship between two factors:

- (1) The rise and fall of a column of mercury in a thermometer with the rise and fall of the outside temperature. As the temperature rises the column of mercury also rises.
- (2) The direction of the wind and the movement of smoke from a chimney. The smoke moves away with the wind.
- (3) The tendency of pupils who are intelligent to be good silent readers.

Negative correlation may be illustrated by:

- (1) The movement of the elevator cage and the counter-balancing weights. As the elevator cage goes up, the counter-balancing weights move in the opposite direction.
- (2) The relation between absence from school and school achievement.

Zero or indifferent correlation is best illustrated by means of the chance matching of numbered cards which have been shuffled. Two packs of 25 blank cards each may be numbered and the packs carefully shuffled so that the cards are in no definite order. If cards are drawn at random from each pack and paired, the resulting relationship is likely to be close to zero. If these same packs of cards are both arranged in ascending order and the first card from one paired with the first card from the other, the resulting relationship is positive. If one pack is inverted and each time a small-numbered card is taken from the one pack a large-numbered

card is taken from the other, the resulting correlation is negative.

This illustration of the numbered cards suggests one of the simple methods of expressing correlation; *viz.*, the method of ranking. If pupils are given two tests and the scores from the tests tend to place the same pupils in the same relative positions in each series, there is an indication of a positive correlation between the two tests. For example, the accompanying pairs of scores earned by nine pupils indicate a high

TABLE XXXVI
PAIRS OF TEST SCORES

Pupil Number	Score on	
	Test A	Test B
1	89	32
2	85	31
3	83	29
4	80	28
5	76	26
6	70	24
7	65	21
8	61	20
9	54	18

positive relationship between the two tests because the pupil making the highest score on Test A also made the highest score on Test B, and each other pupil in the list maintained his relative position. This suggests that the two tests measure abilities which have a great many factors in common. On the other hand, if it had happened that Pupil No. 1, who made a score of 89 on Test A had made a score of 18 on Test B, and Pupil No. 2, who made a score of 85 had made a score of 20 on Test B, etc., the resulting correlation would have been negative, and would have shown that the two tests were inversely related. That is, the negative correlation would indicate that the presence of ability on Test A implied the absence of ability on Test B.

The coefficient of correlation is computed from data arranged in a frequency table, but the table used is a slightly

different form from that employed in any of the problem work in this book thus far. The method of tabulating paired data is explained in the following section.

Tabulating Test Scores in Double Entry or Correlation Tables. The double entry or correlation table is used in computing a coefficient of correlation by the *Pearson Product-Moment Method*. It is also useful frequently when more than one test is given to the same group of subjects. The principles used in setting up the class-intervals in this double entry table are identical with those used in making a grouped frequency table. In fact, the correlation table may be thought of a grouped frequency table on a vertical or y -axis which also has been extended in a horizontal direction to show the place of each frequency in the distribution on the x -axis as well. Thus each tabulation in a correlation table stands for the result of two measures of the same individual. Table XXXVIII on page 535 is an illustration of this type of double entry or correlation table; it is also used to illustrate the steps in the computation of the coefficient of correlation itself. Table XXXVIII shows that one pupil who read with a comprehension score of between 7.5 and 10.5 also read with a rate score of between 59.5 and 66.5 as measured by this particular test, etc. The tendency of the frequencies to group themselves along the diagonal of the table is itself an indication of the correlation which exists between the two factors.

Steps in the Computation of the Pearson Product-Moment Coefficient of Correlation. The calculation of the *Pearson Product-Moment Coefficient of Correlation* involves

the solution of the following formula:
$$r = \frac{\frac{\sum xy}{N} - c_x c_y}{\sigma_x \sigma_y},$$
 in

which r is the coefficient of correlation, N the number of cases, σ_x the standard deviation (in steps) of the distribution on the X -axis, σ_y the standard deviation (in steps) of the distribution on the Y -axis, c_x the correction on the X -axis, c_y the correction on the Y -axis, and $\sum xy$ the sum of the products of the deviations of each measure from the central tendency of the X - and the Y -axes. The solution of the formula involves the following specific steps:

- (1) Tabulation of paired scores in a double entry table.
- (2) Computation of the standard deviation (σ) of each axis of the distribution, leaving σ in terms of *steps*.
- (3) Computation of the sum of the products of the deviations of each measure from the central tendency of both axes (known in the formula as Σxy)
- (4) Division of the Σxy by the number of cases.
- (5) Subtraction of the algebraic product of the corrections of each axis.
- (6) Conversion of the resulting fractional value into decimal form, retaining its sign. If the sign is negative the relationship is inverse and if positive it is direct. The possible values of this decimal range from plus 10 through the possible decimals to 0 and minus 10.

Of these steps (1) and (2) have been explained and illustrated in previous discussion. The major new step is the determination of the sum of the xy products.

The name "product-moment method" implies the significant feature of the process. The relationship itself takes into account the operation of forces (frequencies) at varying distances (deviation in steps) from the point of rotation (the mean). Since each measure assumes a position with regard to both axes the resulting moments⁶ must take this fact into account. For example, if a child makes a high score on each of two tests tabulated in double-entry form, his position will be well up in the upper right quadrant of the table.⁷ The moment represented by his position is, however, the product of the distance his score lies from the mean on the X-axis and the distance his score lies from the mean on the Y-axis as well. In the illustration of Table XXXVII, such a score might be shown by a frequency in the compartment made by the intersection of the deviations plus 4y and plus 3x. The resulting xy product (moment) of such a case is $4 \times 3 \times 1$ or 12. If two cases were found at that point the moment would be 24. The moments or products in the upper right-hand and lower left-hand quadrants are always

⁶ The reader will recall from his high school physics that the term *moment* is given to the force which tends to produce rotation in an unbalanced body. For example, a beam resting on a fulcrum is brought into balance when the moments of the forces operating on it are equal. A weight or a force of two units operating on a beam at a distance of one unit from the point where it is suspended is equalled in moment by a weight or force of one unit at a distance of two units from the point of suspension. Similarly in mathematics, forces are brought into balance by equating their moments.

⁷ Assuming tabulation upward and to the right.

plus if the tabulation is made upward and toward the right. The products in the upper left-hand and lower right-hand quadrants are similarly always negative in sign. The Σxy is the algebraic sum of the plus and minus products. The complete process of computing a product-moment correlation is shown in Table XXXVIII.

Illustration of the Computation of the Pearson Correlation Coefficient. The data for the illustration of Table XXXVIII represent the comprehension and rate scores of seventy-six pupils as measured by a certain silent reading test. The rate scores (the speed of reading) are tabulated along the horizontal (X) axis, and the comprehension scores are tabulated along the vertical (Y) axis. The student should study this illustration until he becomes familiar with the steps involved. Only a very few new processes are introduced. It should be noted that the columns headed d , fd , fd^2 are required in finding the standard deviations and have nothing to do with the values marked xy .

TABLE XXXVII
ILLUSTRATING THE COMPUTATION OF xy PRODUCTS
X-AXIS

Y-AXIS							I		4
									3
			-				+		2
									I
									0
									-I
			+				-		-2
									-3
									-4
	-4	-3	-2	-I	0	I	2	3	4

TABLE XXXVIII

CORRELATION TABLE SHOWING RELATION OF RATE AND COMPREHENSION
AS MEASURED BY A CERTAIN READING TEST

	50 5	60 5	70 5	80 5	90 5	100 5	110 5	120 5	f	d	fd	fd ²	xy
52 5-55 5								2	2	+7	+14	98	42
49 5-52 5									0	+6	0	0	0
46 5-49 5								4	4	+5	+20	100	60
43 5-46 5							1	4	5	+4	+20	80	56
40 5-43 5						4	1	2	7	+3	+21	63	36
37 5-40 5						4	1	3	8	+2	+16	32	30
34 5-37 5					5	2	1		8	+1	+8	8	4
31 5-34 5					5	1		1	7	0			
28 5-31 5				4		3			7	-1	-7	7	8
25 5-28 5			2	4		1	1		8	-2	-16	32	26
22 5-25 5			4	5					9	-3	-27	81	66
19 5-22 5		3	1	1		1			6	-4	-24	96	68
16 5-19 5		1		1		1			3	-5	-15	75	30
13 5-16 5			1						1	-6	-6	36	18
10 5-13 5									0	-7	0	0	0
7 5-10 5	1								1	-8	-8	64	40
f	1	4	8	15	0	16	12	4	16	76			
d	-5	-4	-3	-2	-1	0	+1	+2	+3				
fd	-5	-16	-24	-30	0		+12	+8	+48	(-7)			
fd ²	25	64	72	60	0		12	16	144	(393)			
											(-4)	(772)	(484)

$$e_x = \frac{-75 + 68}{N} = -.092$$

$$e_y = \frac{-103 + 99}{N} = -.053$$

$$\frac{\Sigma xy}{N} = \frac{+484}{76} = +6.37$$

$$S_x^2 = \frac{393}{N} = 5.17$$

$$S_y^2 = \frac{772}{N} = 10.16$$

$$e_x e_y = .005$$

$$\sigma_x \sigma_y = 7.22$$

$$\sigma_x = 2.27$$

$$\sigma_y = 3.18$$

$$r = \frac{\frac{\Sigma xy}{N} - e_x e_y}{\sigma_x \sigma_y} = \frac{6.37 - .005}{2.27 \times 3.18} = +.882$$

There are possibly a few points in the computation of the correlation coefficient which deserve a little further comment. The work of finding the standard deviations (sigmas) of each axis of the double entry table should give no difficulty. *It will be noted, however, that the sigmas in this illustration are not expressed in units of the scale but are left in terms of steps.* The reason for this lies in the fact that, if the sigmas are turned into units of the scale, each of the xy expressions of deviation must also be turned into similar form. This would involve considerable multiplication with a corresponding amount of division later, all of which is useless and only introduces additional opportunities for errors in computation. The value of the coefficient of correlation as a rule is not seriously affected by the size of the step taken in making the table. It is assumed, of course, that the class-intervals will be reasonable, but it is not necessary that they be of the same size on both axes.

Care must be taken to keep the sign of the corrections on each axis. If one correction is positive and the other negative, the resulting $c_x c_y$ product is of course negative. This results in an algebraic subtraction of a negative quantity which serves to increase the numerator of the formula for r and accordingly to increase r itself.

The only other point which need cause the student any trouble is in determining the sum of the xy products ($\sum xy$). In Table XXXVIII, page 535, the frequency 2 which appears in the intersections of the class-intervals 52.5 to 55.5 (Y) and 115.5 to 122.5 (X) lies above the arithmetic mean of the Y -axis seven (7) steps. At the same time these cases lie to the right of (above) the mean of the X -axis three (3) steps. The actual moment of these two cases is therefore the product of the distance they deviate from the mean of the one axis by the distance they deviate from the mean of the other axis ($2 \times +7 \times +3$ or $+42$). The sign of the deviation involved must be carefully noted. Again, two of the eight cases which lie in the class-interval 25.5 to 28.5 on the Y -axis deviate -2 steps from the mean of the one axis and -3 steps from the mean of the other. The product moment of these two cases is therefore $2 \times -2 \times -3$ or $+12$. The product-moment of the 4 cases in the

adjacent cell is $4 \times -2 \times -2$ or $+16$. The one case which deviates -2 steps from the mean of the one axis and 1 from the mean of the X -axis has a product-moment of $1 \times -2 \times +1$ or -2 . The total of the plus moments (12 and 16) is $+28$. The resultant sum of the xy s for this particular row is therefore $28 - 2$ or $+26$. The other xy s in this illustration are found in a similar manner.

From this point in the calculation of the correlation coefficient the work is merely that of substituting values in the formula itself. The sum of the xy products ($+484$) is to be divided by the number of cases in the scatter diagram (76). The result is $+6.37$. From this must be subtracted the product of the corrections⁸ of the X - and Y -axes ($c_x c_y$). This reduces the numerator of the fraction represented by the formula for r to $+6.365$. The product of the sigmas of the X - and Y -axes is 7.22 . The decimal resulting when 6.365 is divided by 7.22 is $+0.882$, the coefficient of correlation (r) of the data in this illustration. In general, this coefficient of correlation indicates that there is a positive and rather significant relationship between the rate of reading of the seventy-six pupils tested and the comprehension with which they read.

Meaning of the Correlation Coefficient. The method of calculating the correlation coefficient as outlined in the foregoing pages is quite mechanical and, as such, can be mastered readily by most students. The interpretation of the meaning or significance of a correlation coefficient is often quite another matter, for no entirely satisfactory mechanical device for accomplishing this has thus far been developed. A number of suggestions have been made recently, however, by means of which the student may be aided in attaching meaning to the correlation coefficient.

It will be remembered that correlation is usually indicated by means of what is called a *double-entry table*, or *correlation table*. The distribution of scores in such a table is known as a "scatter diagram." The appearance of the scatter diagram itself gives some indication of the amount of relationship which exists between the two variables shown. Assuming that the scatter diagram is made by tabulating upward

⁸ This is the same correction used in finding the arithmetic mean.

and toward the right, which is the best uniform practice, a high r is usually found where there is a very definite clustering of the cases along what would be the lower left to upper right diagonal of the table. This means that the cases tend to be grouped somewhat systematically in a practically straight line running from the lower left-hand corner of the table to the upper right-hand corner of the table. This type of grouping is shown in Table XXXVIII on page 535. As the cases scatter from the line of this diagonal, the correlation is reduced. If the cases are scattered over the table in a generally circular arrangement, the resulting correlation will approximate zero. As the relationship changes from positive to negative, the elliptical grouping of the cases takes place along a diagonal running from upper left to lower right. After some experience with "scatter diagrams," the student will come to have a definite feeling as to the probable magnitude of the correlation to expect.

TABLE XXXIX
PERCENTAGE OF FORECASTING
ACCURACY FOR SPECIFIC
VALUES OF r

Coefficient of Correlation	Percent of Forecasting Efficiency
1.00	100
.99	86
.98	80
.95	69
.90	56
.866	50
.80	40
.75	34
.70	29
.65	24
.60	20
.50	13
.40	8
.30	5
.20	2
.10	$\frac{1}{2}$

One of the important outcomes of the use of correlation methods is that within certain limits it makes possible the estimating of unknown values from known values. The accuracy of this estimate, however, depends directly upon the correlation between the factors measured. For example, if it is known from previous experience that there is a high positive relationship between achievement in a specific subject and the pupil's responses to a selected aptitude test, the probable achievement of a group of pupils in this course may be determined within limits by securing their responses to the aptitude test. A correlation coefficient of $+1.0$ for the two factors would mean that an estimate of accomplishment based on the one factor would be 100 percent correct.

As the amount of the correlation decreases, the accuracy of the forecast declines, but not in a direct manner. A correlation of $+1.0$ means 100 percent accuracy in the estimate based on the relationship, but a correlation of $+ .50$ does not mean at all that the estimate based on it will be 50 percent correct. A glance at the accompanying table will demonstrate this interesting fact about the correlation coefficient. The percentages of forecasting accuracy for different values of r given in Table XXXIX are obtained by applying Kelley's formula for the *Coefficient of Alienation* ($k = \sqrt{1 - r^2}$) and then deducting the resulting values expressed as percentages from 100. In cases where estimates of one variable

Value of r	Educational Situation	Interpretation
+ 96	Relation between scores on two forms of a long, analytical reading test for high school pupils.	Evidence of unusually high reliability; scores may be treated with confidence
+ 90	Relation between scores on two forms of a 45-minute group intelligence test.	Evidence of marked reliability.
+ 80	Relation between scores on the same form of a group test of intelligence at the beginning and end of a semester	Evidence of marked relationship; considerable prognostic power even after lapse of time.
+ 50	Relation between scores on a good group intelligence test and course marks of a class in first-year algebra	Evidence of a medium relationship of little value for forecasting purposes (only 13% effective).
- 24	Relation between chronological ages of pupils in a given grade and scores on an objective achievement test.	Evidence of a negative relationship of an indifferent sort, merely shows a slight tendency for the younger pupils in a grade to achieve at a higher level than the average.

are to be made from measurements of another related variable, this table will prove to be a useful safeguard.

A word of warning should possibly be given here in order that the student may not become over-optimistic in the interpretation of correlations. It is better to be on the safe side when making claims for the reliability or the forecasting power of a test. Much irreparable damage has been done to the cause of educational measurements by the unqualified and exaggerated statements of misinformed individuals. As a result of frequent misinterpretation of the measures of relationship, many highly ineffective tests have been accepted and used widely. Many users of educational tests have discovered for themselves the inefficiency of the devices recommended to them and as a result of their experiences will have nothing to do with tests.

The preceding illustrations and practical interpretations of typical correlation coefficients representative of the sort ob-

PROBLEM ON CALCULATING THE CORRELATION COEFFICIENT

Problem 15

Calculating the Pearson Product-Moment Coefficient of Correlation

Prepare a correlation table of the accompanying pairs of scores on the reading and vocabulary sections of a foreign language test, using a class-interval of 3 on each axis. Compute the correlation coefficient ($r = +.867$)

Pupil	Scores		Pupil	Scores		Pupil	Scores	
	Read- ing	Vocab- ulary		Read- ing	Vocab- ulary		Read- ing	Vocab- ulary
1	38	36	11	17	32	21	31	36
2	13	23	12	17	25	22	28	35
3	40	44	13	5	15	23	36	41
4	35	37	14	38	41	24	34	40
5	39	46	15	11	10	25	20	30
6	38	45	16	32	29	26	23	32
7	20	17	17	19	8	27	28	32
8	33	33	18	31	37	28	20	30
9	6	14	19	7	22	29	29	36
10	38	45	20	39	38	30	21	25

tained from educational data have been gleaned from a number of sources. They are offered here for whatever guidance they may give to the student or the teacher in making a critical and conservative interpretation of correlation data.

V. ASSIGNMENT OF RELATIVE AND PERCENTILE RANKS

Need for Method of Ranking. Test scores are given meaning only when they are considered in relation to some standard or norm of achievement. In working with test scores it often becomes desirable to make the achievement of other pupils in the group the basis for the comparison. The relative achievements of pupils may be compared by the simple process of assigning ranks or positions to their scores in accordance with their magnitude. That is, the pupil making the highest score is given first position in the class, the pupil making next highest is given second position, etc. This is called the assignment of *relative ranks*.

The importance of resorting to some device such as the ranking of scores to take care of the fact that scores from different tests or examinations are not comparable may be easily demonstrated. Any teacher readily realizes that a mark of 80 on an examination or a test may mean almost anything. It means one thing when there are no other marks of 80 given in the class. It means quite another thing when there are thirty marks of 80 given in the class. It may also be suggested that it means another thing depending upon how many pupils there are in the class. If the 80 is the highest mark or score made in a class of 50 pupils, such a score is undoubtedly a superior mark. If the 80 is excelled by thirty pupils in the class of fifty, it means that the individual ranks 31st in the class. This point is discussed further in connection with the explanation of absolute ranks.

Assignment of Relative Ranks. *Relative ranks are positions assigned to pupils or scores in terms of the array of which they are a part.* The problem of ranking is merely that of assigning positions to the scores arranged according to size. The largest score is given first position; the next largest, second; etc. The only difficulty involved appears

in the case of two or more scores of the same size. The illustration of Table XL will make clear the method of handling this situation.

TABLE XL
RELATIVE RANKS

Pupil	A	B	C	D	E	F	G	H	I	J	K	L
Score	46	44	44	41	40	38	38	38	37	31	29	28
Rank	1	2.5	2.5	4	5	7	7	7	9	10	11	12

Pupil A, with a score of 46, the highest score when the measures are arranged in descending order of magnitude, is assigned first rank. Pupils B and C, both with scores of 44, would normally be assigned second and third places, but it is impossible fairly to assign second place or third place to one rather than to the other, so the average rank, or a rank of 2.5, is assigned to each. The relative numerical position of each succeeding score is held, however. Pupil D, with a score of 41, being the fourth individual in the series, is assigned fourth place. The position assigned to the pupil with the lowest score should coincide with the number of cases in the series except in the case the last scores are tied.

Care should be taken in the ranking of individuals from scores in order to avoid errors in the assignment of positions. The practice of recording such scores on small cards which can be sorted and arranged in order will eliminate many mistakes if the series runs into many cases.

It must be clear to the student that the assignment of relative rank or position to pupils earning certain test scores actually covers up something of the true situation. For example, the differences in the magnitudes of scores are submerged by the assignment of ranks. To illustrate, in the data used in Table XL a difference of 6 points on the test scale (from 31 to 37) makes a difference of only one rank position (from 9th position to 10th). A difference of only

one test point (from 28 to 29) also makes a difference of one rank position. This is a point which should be kept in mind when using the method of relative ranks. Thus ranking tells us that one pupil is ahead or behind another pupil but fails to tell how much in terms of actual test scores he is ahead or behind the other pupil.

PROBLEMS IN ASSIGNING RELATIVE RANKS

Problem 16

Assigning Relative Ranks

Assign relative ranks to the 27 algebra quiz scores listed for Problem 1, page 502.

Problem 17

Assigning Relative Ranks

Assign relative ranks to the 30 history test scores listed for Problem 2, page 502.

Assignment of Percentile Ranks. The usefulness of the above method is somewhat limited by the fact that it takes no account of the actual level at which the accomplishment takes place. A person ranking 32 in a group of 35 has a very low relative rank in this group. However, if he ranked 32 in a group of 250 the significance of his accomplishment would be greatly changed. *Percentile ranks* take this point into account by reducing the ranking to a basis of 100 units. A percentile score of 75 means that for the measures under consideration the individual assigned that rank made a score which exceeds 75 percent of the individuals of his group without regard to the number of cases it contains.

The student will recognize this 75th percentile as a measure with which he has already had some experience. The 75th percentile of a distribution is exactly the same thing as the third or upper quartile (Q_3). The same methods by which the 25th, 50th, or 75th percentiles are obtained may be applied to the computation of any designated percentile. It is often useful to determine the 10th, 20th, 30th, 40th, etc., percentiles. These are called *deciles*, and are computed in the same manner as the median and the quartiles.

TABLE XLI
COMPUTATION OF DECILES

Class Intervals	<i>f</i>	Cumulative Frequency ^a	Number of Cases to Count Upward	Percentile Scores
70 5-73	5	25	90% of 25 = 22.5	90th = 63.75
67 5-70	5	24	80% of 25 = 20.0	80th = 60.50
64 5-67	5	24	70% of 25 = 17.5	70th = 58.00
61 5-64	5	23	60% of 25 = 15.0	60th = 55.50
58 5-61	5	21	50% of 25 = 12.5	50th = 54.25
55 5-58	5	18	40% of 25 = 10.0	40th = 53.00
52 5-55	5	15	30% of 25 = 7.5	30th = 50.25
49 5-52	5	9	20% of 25 = 5.0	20th = 46.50
46 5-49	5	7	10% of 25 = 2.5	10th = 39.00
43 5-46	5	5		
40 5-43	5	3		
37 5-40	5	3		
34 5-37	5	2		
31 5-34	5	1		
28 5-31	5	1		
<i>N</i> = 25				

Computing Percentiles and Deciles in a Grouped Frequency Distribution. The calculation of the 10th, 20th, 30th, etc., percentiles involves the same general procedure as that followed in computing the median. The only significant difference is, of course, in the number of cases used in counting into the frequency distribution. In computing the median (50th percentile) exactly one-half of the scores are counted off. In computing the 25th percentile (Q_1) one-fourth of the cases are considered. In computing the 10th percentile, quite naturally only one-tenth of the cases are counted off. The interpolation into the proper step is exactly identical with the method followed in computing the median. Table XLI shows the general methods and results without detailed comment. Since the student has already done work of a similar nature, no difficulty in understanding the correctness of every figure in the table should be found.

^a The cumulative frequencies which are obtained by taking the cumulative total of the cases from the bottom of the distribution toward the top are given in this table for convenience in the necessary interpolations.

TABLE XLII
INTERPRETATION OF PERCENTILES

Percentiles (Deciles)	Interpretation	Percentile Scores
100	Score equaled or excelled by no pupil	71 00
90	Score equaled or excelled by 10% of the pupils	63 75
80	Score equaled or excelled by 20% of the pupils	60 50
70	Score equaled or excelled by 30% of the pupils	58.00
60	Score equaled or excelled by 40% of the pupils	55.50
50	Median—score equaled or excelled by 50% of the pupils	54.25
40	Score equaled or excelled by 60% of the pupils	53.00
30	Score equaled or excelled by 70% of the pupils	50 25
20	Score equaled or excelled by 80% of the pupils	46 50
10	Score equaled or excelled by 90% of the pupils	39 00
0	Score equaled or excelled by practically 100% of the cases	30 00

The data of Table XLI give the percentile scores reported in Table XLII. The score assigned as the 100th percentile is the highest score obtained by any pupil on the test used in the table given as an illustration. In this case it is assumed to be 71. The percentile score assigned as zero is the lowest score made by any pupil on the test. In this illustration it is assumed as 30.

The interpretation of percentile scores frequently gives some trouble to the worker inexperienced in their use. Figure 30 is a graphic presentation of the percentile scores given in Table XLII. This figure shows the characteristic curve (ogive) resulting from the use of percentile scores. The heavy solid line in the figure represents the results of an arbitrary smoothing of these percentile scores. This

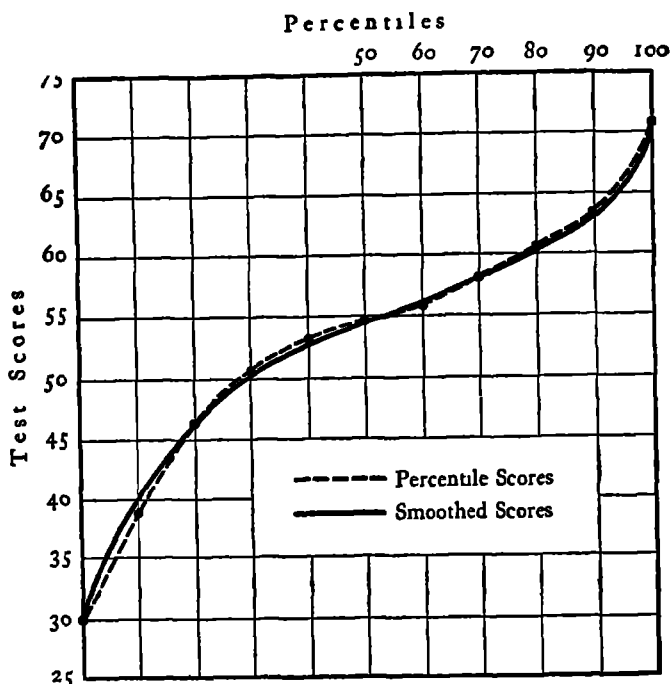


FIGURE 30 PERCENTILE GRAPH

smoothing process is frequently used when percentile scores are based on fairly large populations and are set up as tentative norms for the interpretation of tests. Further reference to norms of this type will be made in Chapter XXIII.

PROBLEMS IN COMPUTING AND GRAPHING PERCENTILE DATA

Problem 18

Computing Percentiles

Compute from the frequency distribution prepared for Problem 2, page 502, the following percentiles 1, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, and 99.

Problem 19

Constructing a Percentile or Ogive Curve

Construct a percentile or ogive curve based on the percentiles computed in Problem 18, using the method outlined in Figure 30 on this page.

VI. SUMMARY

This chapter presents a non-technical discussion of a few of the common statistical tools which the beginning student in educational measurements should find useful in the analysis and interpretation of educational test results. Discussions of four of the six major statistical techniques outlined in the introductory paragraph of this chapter are presented. The fundamental principles of the grouping and tabulating of test scores are stated and illustrated. Methods of expressing the central tendency of measures are presented. The need for measures of dispersion is shown. The quartile deviation and the standard deviation are explained in some detail. The general meaning and the methods of correlation are given, along with a few definite hints concerning the interpretation of correlation coefficients. The practical uses and the meanings of the ranking of test scores are discussed. The two remaining problems, dealing with the derivation and interpretation of test norms, and the use of simple graphic methods of presenting the results of statistical analysis are reserved for treatment in the following chapter.

There has been no attempt to make this chapter a complete discussion of all of the interesting or even useful statistical techniques. To do this would require a volume in itself. As a matter of fact, the brevity of the treatment makes it impossible to present a sufficient number of illustrations to give the inexperienced worker adequate experience with these statistical problems. Real mastery of these technical skills can come only through their repeated and continuous use. Technical facility will not come from a reading of this chapter, but an appreciation of the problems involved and the meaning of the novel notations and vocabulary should come from repeated study of the discussions and illustrations. Without the type of quantitative thinking emphasized in this chapter, the student of education is forced to use vague methods in his appraisals of educational data.

SELECTED REFERENCES

- Broom, M. E., *Educational Measurements in the Elementary School*, Chapters III-IV. New York McGraw-Hill Book Co., Inc., 1939.
- Broom, M. E., *Educational Statistics for Beginning Students*. New York American Book Co., 1936.
- Garrett, Henry E., *Statistics in Psychology and Education* (Second Edition). New York Longmans, Green and Co., 1937.
- Good, Warren R., *The Elements of Statistics*. Ann Arbor, Michigan Ann Arbor Press, 1933.
- Gray, Clarence T., and Votaw, David F., *Statistics Applied to Education and Psychology*. New York The Ronald Press Co., 1939.
- Greene, Harry A., *Work-Book in Educational Measurements*. New York Longmans, Green and Co., 1937.
- Holzinger, Karl J., *Statistical Methods for Students in Education*. Boston Ginn and Co., 1928.
- Kelley, Truman L., *Statistical Method*. New York Macmillan Co., 1923.
- Kramer, Edna E., *Educational Statistics*. New York John Wiley and Sons, Inc., 1935.
- Lang, Albert R., *Modern Methods in Written Examinations*, Chapter XI. Boston Houghton Mifflin Co., 1930.
- Lee, J. Murray, *A Guide to Measurement in Secondary Schools*, Chapters XII-XIII. New York D. Appleton-Century Co., Inc., 1936.
- Lincoln, Edward A., and Workman, Linwood L., *Testing and the Use of Test Results*, Chapter V. New York The Macmillan Co., 1935.
- Lindquist, E. F., *A First Course in Statistics*. Boston Houghton Mifflin Co., 1938.
- Odell, Charles W., *Statistical Method in Education*. New York D. Appleton-Century Co., Inc., 1935.
- Ott, Arthur S., *Statistical Method in Educational Measurement*. Yonkers-on-Hudson, N. Y. World Book Co., 1925.
- Ross, C. C., *Measurement in Today's Schools*, Chapter VIII. New York. Prentice-Hall, Inc., 1941.
- Seder, Margaret, *Introduction to Testing and the Use of Test Results*. New York Educational Records Bureau, July 1940.
- Sorenson, Herbert, *Statistics for Students of Psychology and Education*. New York McGraw-Hill Book Co., Inc., 1936.
- Symonds, Percival M., *Measurement in Secondary Education*, Chapter XII. New York The Macmillan Co., 1928.
- Thurstone, L. L., *The Fundamentals of Statistics*. New York. The Macmillan Co., 1925.
- Tiegs, Ernest W., *Tests and Measurements for Teachers*, Chapter XII. Boston Houghton Mifflin Co., 1931.
- Tiegs, Ernest W., and Crawford, Claude C., *Statistics for Teachers*. Boston Houghton Mifflin Co., 1930.
- Walker, Helen M., *Mathematics Essential for Elementary Statistics*. New York Henry Holt and Co., 1934.
- Williams, J. Harold, *Elementary Statistics*. New York D. C. Heath, 1929.

CHAPTER XXIII

INTERPRETING THE RESULTS OF TESTING

The purpose of this chapter is the consideration of the following important points in the interpretation of test results:

- a.* Meaning of a test score.
- b.* Meaning and types of derived scores.
- c.* Practical uses of the standard deviation.
- d.* Uses of the correlation coefficient for interpreting test results.
- e.* Interpreting achievement test scores by the use of norms.
- f.* Interpreting results of intelligence testing.

I. TEST SCORES

The problems of summarizing test scores and of interpreting the results revealed by these summaries are very closely related and, were it not for the length and detail of the discussion which they require, should probably be considered in a single chapter. The preceding chapter deals with four of the six major problems of statistical procedure as related to test analysis. The content of this chapter concerns itself almost entirely with the last two of these problems of test interpretation.

Meaning of a Test Score. Test scores are valuable to the classroom teacher to the extent that they can be interpreted. It is therefore important to define clearly what is meant by a test score. In order to accomplish this, two or three new concepts require explanation. In the first place, *a test score is a numerical expression of performance on the part of an individual.* Sometimes the test score is merely the number of exercises responded to correctly. Again it may be an arbitrarily defined scale value. But whatever its form, its function is to reveal in a quantitative way the performance of an individual as he responds to stimuli given under certain conditions. This leads to the second concept involved in the meaning of a score. The test score is an evidence of performance. Performance, the response of the individual to the test situation, is taken to mean in educational measurements the expression of ability operating under

certain conditions. *Performance may be thought of as Ability + Conditions.* Scores on tests are definitely influenced by conditions. The pupil may make a poor score because he does not have the ability to do better—may not know the facts, etc. On the other hand, he may make a low score because of illness, discomfort, hearing, sight, illumination, a broken pencil, indifference for the subject, dislike for the teacher or examiner, failure to give attention to and to comprehend the directions, etc. Any one of these or a dozen other factors may affect the score. Accordingly, there is the possibility and even the likelihood of a serious error in the assumption that a test score is a direct evidence of ability. The conditions under which the performance takes place must be known before it is safe to infer ability from performance.

Ability, as an abstract concept, may be defined as the power to do. Power to do, to respond to stimuli and to situations, is the product of training and experience. *Ability may be thought of as Capacity + Training*, which suggests that unless training and native capacity factors are known, inferences as to abilities may be misleading. This point becomes particularly serious in the interpretation of intelligence test results, for it is a common practice for users of such tests to infer innate capacity (mental ability) from performance scores. The real seriousness of this type of uncritical inference may be seen by comparing the interpretations of an achievement test score and an intelligence test score. Both are basically expressions of performance. Equal abilities may be inferred from equal scores in both types of tests if and when the conditions under which they are given are all definitely under control. While it is difficult to make sure that all physical and physiological factors are adequately controlled in a testing situation, it is possible to regulate most of the mechanical conditions within reasonable limits. The significant point to note here, however, is the fact that users of achievement tests stop with an inference of equality of ability from equal performance scores, but users of intelligence tests are obliged to make a further inference. In the interpretation of intelligence test results, it is common practice to infer equal native capacity from apparent evidences of

equal abilities. The fallacies in this argument and the dangers of this step must be readily apparent. *Equal capacities may be inferred from performance scores only when there is direct and positive evidence of two things: first that the conditions under which the testing took place were identical and equally well controlled; second, that the training opportunities of the individuals compared have been equal.* The mechanics of testing now make it fairly easy to control testing conditions. The second factor represents a real stumbling block in the way of an accurate and sane interpretation of intelligence test results. The naive manner in which some makers and many users of such tests assume equality of learning opportunity, and hence equal capacity from equal performance scores, is one of the things which has made many teachers and students skeptical of their value.

The foregoing discussion of the meaning of a test score may appear to indicate that it is impossible to give meaning to any kind of test score. Such is not the intention, even though the purpose here is to emphasize the need for a conservative attitude in test score interpretation. In the long run, the more that is known about the variables underlying test scores, the more critical must the user become. The greatest damage that has been done to the field of educational measurements has come as a direct result of carelessness and ignorance on the part of users of tests, and their tendency to draw unwarranted conclusions from the results. The teacher should be able critically to select suitable tests and scales for classroom use, to control the mechanical conditions of their administration, and to draw sane and defensible conclusions and inferences from the results.

Giving Meaning to Informal Test Scores. The user of educational tests in the classroom is confronted with two types of test data for interpretation. The first type, and undoubtedly the more common of the two, deals with the results of informal, teacher-made tests. The results from these home-made tests are in turn of two types—the subjective scores assigned by teachers to pupils' responses to essay-type tests, and the performance scores resulting from informal objective examinations. While something can be done to improve the interpretation of the relatively unre-

liable marks assigned to the discussion-type exercise, much more is possible in the complete and accurate interpretation of the performance scores which are the results of long and reliable objective examinations. Since one of the major functions of the standardization of a test is the establishment of meaning for the test scores, many additional types of interpretation are thus made possible in the use of the second type of test data, which is obtained from standardized tests.

II. TYPES OF DERIVED SCORES

Function of Derived Scores. Test scores which are used to describe the performance of pupils are expressed in a variety of different units and in relation to a variety of different scales of measurement. In some tests the scale of measurement or the unit of measurement used may be relatively large. Pupil scores expressed in terms of these large units may be small. A long test composed of many exercises may result in large scores. Some common basis must be utilized if comparisons of scores based on these widely different types of scales are to be possible.

This lack of comparability between raw scores from different tests is illustrated by the fact that two different well-known reading tests for Grades 3 to 8 have median scores for sixth-grade pupils of 13 and 52. Similarly, median performances for pupils sixteen years of age are represented by scores on two popular intelligence tests of 61 and 133. It is thus apparent that a given point score on one test may mean exceptional achievement while a similar point score on another test may mean exceedingly poor achievement for the grade. To repeat, if scores on different tests are to be compared directly, they must be reduced to a common denominator. That is, if a score of 62 points on one test is poor performance, but on a second test because of a different system of scoring is very good performance, some way of translating both of these scores of 62 points to a common system of scores must be devised so that their difference in meaning will be clear. A number of proposals have been made for the calculation of derived scores which will partially take care of this difficulty.

Relation of Derived Scores to Norms. Confusion may easily exist in the thinking of the student concerning the distinction between derived scores and norms. As a matter of fact, tables of norms often yield such derived scores as grade scores, age scores, or percentiles. The use of norm tables for obtaining such derived scores directly from raw scores or point scores on various tests is illustrated in Chapter V and is also discussed later in this chapter. However, such ratios as the intelligence quotient, educational quotient, accomplishment quotient, reading quotient, etc., are derived scores, but they are obtained by a division of one value by another. Although tables for determining such quotients are available, they are not tables of norms but tables to facilitate certain arithmetical computations. Some of these quotients were discussed in previous chapters, and will also be presented in another manner later in this chapter.

Another possible source of confusion to the student lies in the fact that some tests provide a two-step procedure from raw scores to norms. In such cases, such derived scores as standard scores, scaled scores, converted scores, C-scores, etc., are obtained from raw scores. Such derived scores have more meaning than do raw scores, but they do not always have final meaning for the interpretation of test results. Consequently, it is necessary to enter a table of norms with the derived scores and to interpret them in terms of such other derived scores as grade scores, age scores, or percentile scores. Situations of this type will also be discussed in a later section of this chapter.

It is believed that the most satisfactory method of familiarizing the student with derived scores and norms is first to present the various types of derived scores, methods of computing them, and something of their meaning, and then to illustrate the need for and use of norms in the further interpretations required to make some of them meaningful. In the treatment which follows, three types of derived scores are distinguished: (1) those based on average or median performance, (2) quotients and related measures, and (3) those based on variability of performance.

Derived Scores Based on Average or Median Performance. The two types of derived scores which are based on

average or median performance are the grade score and the age score. These two types of derived scores are directly dependent upon tables of norms, for it is only by entering norm tables with raw scores or some other forms of scores that grade equivalents or age equivalents can be determined. The meaning of grade scores and age scores is presented here and the use of tables of norms for their derivation is illustrated later in this chapter.

Grade Equivalents as Derived Scores. A grade equivalent indicates the position on a grade scale at which a pupil's test performance places him. For example, a child may attain a score on a reading test which is identical with the average or median score of pupils three months into the fourth grade. If so, his grade equivalent on the subject matter of the test is 4^3 , regardless of whether he may be in the fourth grade or in some grade above or below the fourth. Grade scores are sometimes referred to as G-scores or as B-scores. Grade and months are commonly listed as a number and its exponent respectively or as a number and a decimal respectively. Thus the above grade equivalent might be stated either as 4^3 or as 4.3.

Age Equivalents as Derived Scores. In a manner very similar to that which operates for grade scores, age equivalents indicate the position on an age scale at which a pupil's test-performance places him. The hypothetical child whose reading test score gave him a grade equivalent of 4^3 , for example, might be found by the use of a table of age norms to have an age equivalent of nine years eight months (9-8) on the same test. This would mean that his score was identical with the score made by the average or median child nine years and eight months of age. He might actually be a year or so older or younger; his age equivalent on the subject matter of the test would nevertheless be 9-8. Age equivalents are represented by such terms as educational age (EA) for achievement over broad areas of subject matter, mental age (MA) for performance on general intelligence tests, and reading age (RA) for achievement in reading skills. Such ages are commonly stated in hyphenated form, the first number indicating years and the second number months of age. Thus the EA of 9-8 indicates that in broad

educational achievement the child used in the above illustration is at the same level as average children nine years and eight months of age.

Although this book is most directly concerned with the types of age equivalents noted above, the same technique is applied to the measurement of other aspects of child growth and performance. For example, anatomical age, physiological age, and social age are comparable terms which are employed with varying degrees of exactness in meaning. Chronological or life age is, of course, the most widely used of all, and is frequently employed as the basic or criterion measure of test validity, as will be pointed out in the following paragraphs.

Quotients as Derived Scores. Quotients and other similar derived scores show the relationship existing between two characteristics for the child as a means of indicating the manner in which growth of various types is related. For instance, the educational quotient, intelligence quotient, and reading quotient are ratios respectively between a child's educational or mental and chronological ages. The accomplishment quotient is the ratio between a child's educational and mental ages. The first three are based on the idea that on the average a child grows in all ways more nearly in conformance with his chronological age than with any other measures, and also upon the recognition that deviations from that pattern of growth result from individual differences and are meaningful in the guidance of the child. The accomplishment or achievement quotient is based on the idea that the child's mental age is perhaps a better criterion by which to judge his educational growth than is his chronological age. All of these have been discussed in appropriate chapters elsewhere in this volume.

Computation of the various quotients listed above will be illustrated for a pupil who has, say, the following ages: (1) chronological age (CA) of 8-4, (2) educational age (EA) of 9-2, (3) mental age (MA) of 9-7, and (4) reading age (RA) of 9-4. The last three ages would be determined in the manner indicated in the above section from his scores on a general achievement, a general intelligence, and a reading test. The quotients are all based on computations in which

each age is reduced to months, and all ratios are multiplied by 100 to eliminate the use of decimals in the results.

For the child whose various age levels or age equivalents are given above, his educational quotient (EQ) would be

$$EQ = 100 \frac{EA}{CA} = 100 \frac{9-2}{8-4} = 100 \frac{110(\text{months})}{100(\text{months})} = 110,$$

his intelligence quotient (IQ) would be

$$IQ = 100 \frac{MA}{CA} = 100 \frac{9-7}{8-4} = 100 \frac{115(\text{months})}{100(\text{months})} = 115,$$

his reading quotient (RQ) would be

$$RQ = 100 \frac{RA}{CA} = 100 \frac{9-4}{8-4} = 100 \frac{112(\text{months})}{100(\text{months})} = 112,$$

and his accomplishment quotient (AQ) would be

$$AQ = 100 \frac{EQ}{IQ} = 100 \frac{110}{115} = 95.6 \text{ or } 96,$$

or

$$AQ = 100 \frac{EA}{MA} = 100 \frac{9-2}{9-7} = 100 \frac{110}{115} = 95.6 \text{ or } 96.$$

These quotients indicate that the child is well above average for his age in intelligence and is somewhat less accelerated educationally. Within the limits of reliability for the AQ, discussed in some detail in Chapter X, it appears that his achievement is not quite what might be expected of a child of his mental ability level. His reading quotient indicates a somewhat greater advancement in that subject than for the average of all other areas of achievement covered by the general achievement test from which his EA was determined.

With this brief presentation of the method of deriving the various commonly used quotients as a background, the student should be able to interpret these quotients adequately when he encounters them elsewhere in this volume. It should be understood that the RQ is merely representative of quotients which can be derived for the various subjects of the curriculum if age norms are given for such subjects on the standardized tests which are used. In prac-

tice, such quotients are seldom used except for reading and arithmetic, however.

Derived Scores Based on Variability of Performance. Derived scores which are based on variability of performance are of two types: (1) percentile ranks, and (2) scores which express position on a scale in terms of the standard deviation or quartile deviation as a measure of variability. Although these two methods are similar in some respects, they differ in several fundamentally important ways which determine their relative effectiveness for certain types of uses. Percentile scores are less reliable than are derived scores based on the *S.D.*, because percentiles are much more affected by minor fluctuations in the distribution of scores upon which they are based than is the standard deviation. Percentile scores cannot with strict validity be averaged, whereas averaging of several scores similarly stated in terms of the *S.D.* is a defensible procedure. Percentiles are based on equivalent areas under the distribution curve, so that percentiles of, say, 48 and 49 usually represent closely similar scores, whereas percentiles of 2 and 3 may represent scores a number of raw-score units apart. On the other hand, derived scores based on the *S.D.* or *Q* differ by equivalent distances along the scale, so that they represent merely the application of a new and more meaningful linear scale to a linear distance.¹

Percentile Ranks as Derived Scores. The test performance of a pupil may be expressed in terms of the position which his score occupies in the standard distribution of scores for pupils (1) of a particular grade, (2) of a certain course, such as plane geometry, or (3) having studied a certain subject, such as a foreign language, for a given number of semesters. This is accomplished by dividing the distribution so that exact divisions contain the same percentage of the total number of cases. Various plans are to divide the distribution into quarters (quartiles), fifths (quintiles), tenths (deciles), or hundredths (percentiles or centiles). When the distribution is divided into one hundred parts, each part is most commonly called a percentile. Ability as rep-

¹ Interested students can locate more complete discussions of these differences in practically any of the standard statistics books listed at the end of this chapter.

resented by a test score is indicated by a number between 0 and 100 which expresses the percentage of this standard group falling below the particular score in question. Percentiles are particularly useful with high school tests, for which grade and age norms are not ordinarily meaningful. An illustration of the method of obtaining deciles from a distribution of scores was given in the preceding chapter.

Derived Scores Based on the Standard Deviation or Quartile Deviation. A considerable number of derived scores have the standard deviation and arithmetic mean of a standard group of pupils as basic to their derivation. These various derived scores have different names, and some of them are devised for use with particular tests or series of tests. Although they differ widely in the manner in which the standard groups upon which they are based are selected, and make use of different numerical methods of representation, they have the element in common of being based upon the standard deviation.

The arithmetic mean and standard deviation were both presented in the preceding chapter. One of their major uses is found in their provision of one of the most satisfactory means of deriving meaningful scores from test results. The brief treatment of derived scores here shows the major types of such scores and the elements of similarity and difference among them.

Standard Measures or z-scores are mentioned briefly here because they represent such a simple method of showing deviation of a score from the arithmetic mean of the distribution and because of their similarity to other derived scores. However, the z-score is a measure used primarily in statistical procedures, and has very little direct significance for the interpretation of test results to the teacher. The z-score is found by the application of the formula

$$z = \frac{X - M}{S.D.},$$

in which X is a particular raw score, M is the arithmetic mean of the distribution of raw scores, and $S.D.$ is the standard deviation of the distribution of raw scores. It is sufficient here to point out that the z-score expresses devia-

tion from the arithmetic mean in terms of standard deviation units and to give a few illustrations. For example, a z-score of $+2.00$ is two sigmas above the mean, a z-score of -2.00 is two sigmas below the mean, and a z-score of $-.37$ is $.37$ S.D. below the mean. Therefore deviations from the mean can be read directly from z-scores.

T-scores are similar to z-scores, except that they eliminate the use of negative values and decimals. A T-score of 50 was arbitrarily decided upon to represent a score at the arithmetic mean of a distribution and 10 T-score units were made equivalent to one standard deviation of distance. The formula for the T-score is

$$T = \frac{10(X - M)}{S.D.} + 50,$$

where X , M , and $S.D.$ have exactly the same significance as they had in computing z-scores, that is, a particular raw score, the arithmetic mean, and the standard deviation. A score two sigmas above the mean has a T-score value of 70, a score two sigmas below the mean has a T-score value of 30, and a score $.37$ S.D. below the mean has a T-score equivalent of 46. Fractional values are not ordinarily used in T-scores.

Standard Scores, *Scaled Scores*, and *Converted Scores* are other types of derived scores which provide for comparability of scores on different parts of the same test or even on different tests. This is accomplished by changing raw scores to derived scores by methods differing somewhat from those described above but nevertheless based on the mean and standard deviation for some standard group.

C-Scores are similar to the T-score and scaled score in numerical representation but the C-score unit is one-tenth of a quartile deviation rather than of a standard deviation.

Other Types of Derived Scores. Although the types of derived scores discussed above are those most commonly used, several miscellaneous types which do not fit into any of the categories above merit brief mention here.

In the field of intelligence testing, the personal constant (PC) and the index of brightness (IB) are not mentioned above, but they are given adequate treatment in Chapter X.

Personality inventories in a few instances make use of the personality quotient (PQ), which is treated sufficiently in Chapter XI. Two derived scores which relate intelligence and achievement—the mental and educational indices, and the index of studiousness—are given attention in Chapter X.

The derived scores discussed in this chapter and elsewhere in the volume probably do not include all of the types or variations of such measures, for it is not uncommon to find that a new test appears with a new type of derived score. However, the types presented are the most widely used and the most important at the present time.

PROBLEM IN COMPUTING T-SCORES

PROBLEM 20

Computing T-Scores

Use the results of your work on Problem 13, page 528, and assign T-score equivalents to the 30 history test scores tabulated in Problem 2, page 502

III. PRACTICAL USES OF THE STANDARD DEVIATION

One of the major uses of the standard deviation, as the measure basic to certain important types of derived scores, is treated in an earlier section of this chapter. The standard deviation has many other uses, however, and two of them are important enough to justify attention here.

Assignment of Class Marks. The student or the teacher who is interested in the critical analysis of test scores will find the standard deviation a very useful and reliable instrument for the purpose. For example, it offers the basis for an objective plan for turning scores on objective tests into class marks. The importance of this practice is so great that the steps involved in the technique are given in detail. The computations described are based upon the objective test scores from a class of 45 pupils given in Table XLIII. The student will do well to check all of these computations for errors.

Step 1. Prepare a suitable frequency table of the test scores, lay off the deviations from the assumed mean, find the sum of the fd values, and determine the arithmetic mean. The mean of this distribution is 68.55.

Step 2. Compute the standard deviation (σ) of this distribution in exactly the same way as is illustrated on page 562.

Step 3. Since a distance of two and one-half sigma units above and below the mean includes almost 99 percent of all cases in a distribution, lay off this number of sigma units above and below the mean to mark off the limits for the five marks of *A*, *B*, *C*, *D*, and *Fd*. This naturally results in placing one of the sigma units in the middle of the distribution in such a way that one-half of the sigma distance of the middle unit extends above and one-half below the mean. Accordingly, to the arithmetic mean of 68.55 add one-half of the standard deviation (one-half of 19.40). This gives a value of 78.25, which becomes the upper limit of the group of scores which will be assigned marks of *C*.

Step 4. Find the upper limit of the group of scores to be assigned *B* marks by adding one and one-half standard deviation units to the arithmetic mean. Thus, $68.55 + 1.5 (19.40) = 97.65$, which is the upper limit of the *B* group.

Step 5. Find the upper limit of the *D* group by subtracting one-half of a standard deviation unit from the mean. $[68.55 - .5 (19.40) = 58.85.]$

Step 6. Find the upper limit of the *Fd* group by subtracting one and one-half sigma units from the mean of the distribution. $[68.55 - 1.5 (19.40) = 39.45.]$

Step 7. From these values find the score limits of the five divisions of this distribution. Class marks may then be assigned as indicated to the scores within the limits specified.

Marks	Score Limits
<i>A</i>	97 65 and above
<i>B</i>	78 25 to 97 65
<i>C</i>	58 85 to 78 25
<i>D</i>	39.45 to 58 85
<i>Fd</i>	Below 39 45

It is readily apparent that practically no subjective factors are involved in the assignment of marks by this method. The objective test scores of the forty-five pupils used in the illustration are changed by this treatment into 5*A*, 8*B*, 16*C*, 14*D*, and 2*Fd* marks. The score limits are determined by the standard deviation units and would be the same no matter who assigned the marks. It should be noted,

TABLE XLIII

STANDARD DEVIATION TECHNIQUE FOR ASSIGNING CLASS MARKS

Test Scores		Mid-Points	Class Intervals	f	d	fd	fd^2
109	A (11.1%)	110	107 5-112 5	1	10	10	100
104		105	102 5-107 5	2	9	18	162
103		100	97 5-102 5	2	8	16	128
102		95	92 5-97 5	4	7	28	196
99		90	87 5-92 5	0	6	0	0
95	B (17.8%)	85	82 5-87 5	2	5	10	50
95		80	77 5-82 5	2	4	8	32
94		75	72 5-77 5	3	3	9	27
93		70	67 5-72 5	3	2	6	12
84		65	62 5-67 5	4	1	4	4
83		60	57 5-62 5	7		(+109)	
79		55	52 5-57 5	7	-1	-7	7
79		50	47 5-52 5	4	-2	-8	16
77	C (35.5%)	45	42 5-47 5	1	-3	-3	9
76		40	37 5-42 5	1	-4	-4	16
76		35	32 5-37 5	2	-5	-10	50
71			$N = 45$			(-32)77	809
71							
69							
64							
64							
64	D (31.1%)						
62							
60							
60							
59							
59							
58							
57							
57	E (4.5%)						
56							
56							
55							
55							
53							
52							
51							
51	F (4.5%)						
47							
47							
37							

$$\begin{aligned}
 A.M. &= 60 + s \frac{\sum fd}{n} & S.D. &= s \sqrt{\frac{\sum fd^2}{n} - c^2} \\
 &= 60 + 5 \frac{77}{45} & &= 5 \sqrt{\frac{809}{45} - (1.71)^2} \\
 &= 60 + 5 (1.71) & &= 5 \sqrt{17.98 - 2.92} \\
 &= 60 + 8.55 & &= 5 \sqrt{15.06} \\
 &= 68.55 & &= 5 \times 3.88 \text{ or } 19.40
 \end{aligned}$$

Find score limits:

$$\begin{aligned}
 68.55 + \frac{1}{2} (19.40) &= 78.25 \text{ upper limit of C group} \\
 68.55 + 1\frac{1}{2} (19.40) &= 97.65 \text{ upper limit of B group} \\
 68.55 - \frac{1}{2} (19.40) &= 58.85 \text{ upper limit of D group} \\
 68.55 - 1\frac{1}{2} (19.40) &= 39.45 \text{ upper limit of F group} \\
 A &= \text{Above } 97.65 & D &= 39.45 \text{ to } 58.85 \\
 B &= 78.25 \text{ to } 97.65 & F &= \text{Below } 39.45 \\
 C &= 58.85 \text{ to } 78.25
 \end{aligned}$$

however, that these limits hold only for this particular distribution and must not be assumed to be true for any other test. The teacher should also remember that this method of marking does not take into account the absolute level of ability at which a particular class works. The superior pupil in an average or poor class receives an *A* by this method just as readily as does the superior pupil in a very superior class. This is probably less serious than it sounds, however, for most class groups large enough to warrant the application of this technique average out quite well in this respect.

PROBLEM IN ASSIGNING MARKS FOR TEST SCORES

PROBLEM 21

Assigning Marks for Test Scores

Assign the proper letter mark—A, B, C, D, or F—to each of the 30 history test scores tabulated in Problem 2, page 502, using the method for assigning marks which is outlined above.

Scaling of Test Items. The standard deviation, along with certain other measures of variability, represents a convenient unit in which to evaluate the difficulty of test items. When used under these conditions, the standard deviation of a theoretically normal curve of the specified ability is used as the unit in laying off differences in difficulty along a linear scale. As a first step in the procedure, the percentage of pupils failing on each item or exercise must be secured. By means of tables based upon the normal curve, these percentages of failure are changed into standard deviation units which express the positions of each of the exercises with respect to the mean ability of an infinite and normal population. Exercises which are answered successfully by 50 percent of the class are assigned a position at the mean. Exercises missed by 55 or 60 percent of the class are given sigma values above the mean, etc. A significant feature of this procedure, however, is the fact that a difference in difficulty of five percent near the mean results in a relatively small sigma difference, while a five percent difference near the extremes of the distribution makes a relatively large sigma difference. This is in conformity with the fact that because of the height of the curve near the mean a smaller distance

along the linear scale on the base line is required to add a given area of the curve. Thus, the difference in the sigma values assigned to two test items having percentages of failure of 55 and 60 is .13 standard deviation units, ($2.74 - 2.61$), while the difference in apparent difficulty of two items failed by 90 percent and 95 percent of an experimental group is .34 standard deviation units ($4.09 - 3.75$). The net result of this method of item evaluation is to magnify somewhat the simplicity of the very easy item and the difficulty of the very hard one.

Sigma units are also utilized in the construction of scales for the estimation of the merit or quality of certain classroom products. The use of these units in the derivation of such scales is too limited to warrant treatment in this book.

IV. PRACTICAL USES OF THE CORRELATION COEFFICIENT

The classroom teacher and the student of measurement will find the greatest opportunity to use correlation techniques in connection with the construction and analysis of objective tests and in the critical selection of standardized tests. The uses briefly mentioned and in two cases illustrated below all relate to the determination of test validity, reliability, or objectivity.

Determination of Test Validity. Test validity can be determined in terms of correlations between scores on the test and. (1) teachers' marks, (2) ratings of expert judges, (3) other known measures, and (4) measures of future outcomes. As all of these situations involve only the ordinary application of the correlation method, so, as their values are discussed in Chapter IV, they are not discussed further here.

Evaluation of Test Reliability. The correlation coefficient enters directly into the procedures most common for determining or estimating the reliability or consistency of a test, or the degree to which it measures whatever it does measure. As is pointed out more in detail in Chapter IV, there are three correlation methods and one non-correlation method which can effectively be used by the teacher in estimating the reliability of his classroom tests. These are the (1) reliability coefficient, (2) retesting coefficient, (3)

"chance-half" coefficient, and (4) "footrule" coefficient.

Reliability and Retesting Coefficients require only brief mention here, because they involve only the usual type of correlational relationship between two series of scores. The reliability coefficient itself is obtained only by correlating scores made by the same pupils on two equivalent forms of the same test. The retesting coefficient, requiring correlation of scores obtained from a first and a second administration of the same test to a group of pupils, furnishes an estimate of test reliability. The retesting coefficient is one of the methods used when the availability of only one form of the test eliminates the possibility of obtaining a reliability coefficient directly.

The "Chance-Half" Coefficient is a second method of estimating the reliability coefficient from the results of the administration of a single test to a pupil group. For this method, the first step of procedure is to obtain two "half-scores" for each pupil on arbitrary halves of the test. The arbitrary halves of the test frequently consist of the odd-numbered and the even-numbered items. The second step is to obtain the coefficient of correlation between the sets of half-scores for the group of pupils. This coefficient represents the reliability of *one half* of the test, but not of the entire test.

The third and final step requires the use of the *Spearman-Brown Prophecy Formula* in estimating the reliability for the entire test by what is known as "stepping up" the correlation. As a test increases in reliability as it is increased in length by additional test items comparable to those in the initial test, the estimated reliability for the entire test is greater than for that of only half of the test. However, the increase in the coefficient is not directly proportional to the increase in test length. The Spearman-Brown formula is

$$r_{12} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}},$$

where $r_{\frac{1}{2}\frac{1}{2}}$ is the correlation between scores on the "chance-halves" of the test and r_{12} is the estimated reliability for the entire test.²

² The general form of the formula, which is not of direct concern here, is

If an estimate of the reliability of an entire test is desired when the correlation coefficient between its "chance-halves" is .85, the following result is obtained.

$$r_{12} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}} = \frac{2 \times .85}{1 + .85} = \frac{1.70}{1.85} = .92.$$

This is the procedure a teacher may use to obtain an estimate of the reliability of his test from a single administration.

The "Footrule" Coefficient is a simple method of obtaining an estimate of the reliability of a test available in only one form.³ The only values required are the arithmetic mean and standard deviation of the test scores and the number of items in the test. The formula is

$$r_{tt} = \frac{n}{n-1} \times \frac{\sigma^2 - n\bar{p}\bar{q}}{\sigma^2},$$

where $\bar{p} = \frac{M_t}{n}$ and $\bar{q} = 1.00 - \bar{p}$, and where M is the arithmetic mean of the test scores, σ is the standard deviation of the test scores, and n is the number of test items.

The "Footrule" coefficient of a test of 249 items for which the arithmetic mean and standard deviation of scores were respectively 168.65 and 25.34 would be obtained by the following procedures

$$\begin{aligned} \bar{p} &= \frac{M_t}{n} = \frac{168.65}{249} = .677 & \bar{q} &= 1.00 - \bar{p} = .323 \\ r_{tt} &= \frac{n}{n-1} \times \frac{\sigma^2 - n\bar{p}\bar{q}}{\sigma^2} \\ &= \frac{249}{248} \times \frac{25.34^2 - 249 \times .677 \times .323}{25.34^2} \\ &= 1.004 \times \frac{642.116 - 54.45}{642.116} = 1.004 \times .915 = .919 \end{aligned}$$

$$r_n = \frac{nr_{12}}{1 + (n-1)r_{12}}$$

in which r_{12} represents the coefficient of reliability of a test and r_n represents the coefficient of reliability of a test of homogeneous test materials n times as long. It should be noted that substitution of 2 for n in this formula, to determine the effect of doubling the length of the test, results in the special formula given above except for differences in the subscript for r .

³G F Kuder and M W Richardson, "The Theory of the Estimation of Test Reliability," (Formula 21) *Psychometrika*, 2 151-60, September 1937.

Determination of Test Objectivity. When a group of test papers has independently been scored twice, either by the same person or by different persons, the correlation coefficient between the two sets of scores is the objectivity coefficient. For a highly objective test, the coefficient should closely approach $+1.00$.

PROBLEMS IN ESTIMATING TEST RELIABILITY

PROBLEM 22

Estimating Test Reliability by the "Chance-Half" Correlation Method

Prepare a correlation table of the accompanying pairs of scores representing scores on the odd- and even-numbered exercises in a 100-item objective examination. Use a step of 3 on both axes. Compute the correlation coefficient of these odd-even measures ($r_{\frac{1}{2}\frac{1}{2}} = .956$). Then estimate the reliability coefficient to the entire test by using the *Spearman-Brown Formula* ($r_{12} = .978$).

Pupil	Scores on		Pupil	Scores on		Pupil	Scores on	
	Odds	Evens		Odds	Evens		Odds	Evens
1	47	45	13	29	28	24	20	17
2	45	45	14	26	31	25	19	21
3	44	42	15	26	30	26	17	17
4	41	44	16	26	26	27	15	18
5	40	38	17	25	28	28	15	13
6	37	35	18	24	25	29	14	12
7	35	43	19	23	24	30	12	16
8	35	32	20	23	20	31	12	14
9	34	40	21	23	17	32	10	13
10	34	26	22	20	25	33	8	3
11	31	37	23	20	19	34	6	10
12	30	33				35	2	6

PROBLEM 23

Estimating Test Reliability by the "Footrule" Method

Compute the "Footrule" coefficient for a test consisting of 120 items on which the arithmetic mean and standard deviation of the scores made by a class of pupils were respectively 79.80 and 15.70. ($r_{tt} = +.898$).

V. THE USE OF NORMS FOR INTERPRETING TEST RESULTS

In the interpretation of results from standard tests, tables of norms nearly always have direct value. Norms of one of the three types discussed in Chapter V are ordinarily provided with such tests. (1) grade norms, (2) age norms, or (3) percentile norms. As has been pointed out in a previous section of this chapter, derived scores and norms overlap in various ways. In the section of this chapter devoted to derived scores, the various types of meaningful scores which can be derived either from tables of norms for standardized tests or from statistical manipulations of scores from standardized and informal objective tests were treated. Only those derived scores enter into the discussion of this section which are related to norms in one of two ways (1) as results from the use of norm tables—grade scores, age scores, and percentile ranks, or (2) as scores intermediate between raw scores and final derived scores, e.g., standard scores, equated scores.

The tremendous variety of methods by which norm tables are presented for different standardized tests makes impossible a presentation here of more than a few illustrations, and they cannot be considered to represent all of the variations found in the details of organization of norm tables. Furthermore, the purpose here is only to familiarize the student sufficiently with the nature, form, and use of norms that he will be able to employ norms properly in the interpretation of results from any standardized test he may have occasion to use. By giving intelligent attention to instructions in the manual of directions accompanying most tests, the teacher should encounter few difficulties in test interpretation.

Interpretation of Reading Test Results. An illustration of the interpretation of results in terms of grade norms for the *Thorndike-McCall Reading Scale* is based on the norm data of Table XLIV. Grade equivalents can be found for scores on any test having reliable grade norms. Their significance is, of course, subject to the reliability of the test norms as well as to the reliability of the test itself.

TABLE XLIV
G-SCORES FOR THORNDIKE-McCALL READING SCALES⁴

Crude Score	G—Score	Crude Score	G—Score	Crude Score	G—Score
35	15 6	23	6 5	11	3 3
34	15 3	22	6 1	10	3 1
33	15 0	21	5 8	9	2 9
32	14 7	20	5 4	8	2 9
31	14 4	19	5 1	7	2 8
30	13 3	18	4 7	6	2.6
29	11 7	17	4 5	5	2 5
28	9 2	16	4 2	4	2.4
27	8 3	15	4 0	3	2 3
26	7 7	14	3 8	2	2 1
25	7 2	13	3.7	1	1 8
24	6 8	12	3 5	0	1 5

Table XLIV shows that a crude score of 15 has a grade equivalent (G-score) of 4 0, which represents typical beginning fourth-grade achievement. Each point value from 9-10 to 15 is assigned its proportionate position within the third grade on the scale. A point score of 12 is considered as equivalent to the 3.5 grade, 14 to 3.8 grade, etc. In other words, a score of 12 is such a score as may be expected from average pupils halfway through the third grade.

Grade scores of this type are easily obtained from the grade norms ordinarily accompanying standard tests. The degree of fineness with which they may be obtained depends upon the characteristics of the test scale itself. If the test scale consists of many fine units, the grade scores assigned can be correspondingly detailed. A test with a relatively coarse unit requiring only one or two such units to span the entire difference between two grades would, of course, permit the establishment of the grade scores only on a similarly crude basis.

This method of expressing test scores is particularly useful

⁴ *Directions for Using Form 1, Thorndike-McCall Reading Scale*, p. 2 Bureau of Publications, Teachers College, Columbia University, New York, 1931.

where test results are used for purposes of gradation and classification of pupils. The chief convenience lies in the fact that grade scores express the achievement of the pupil in terms of the school unit of classification.

Use of a Graphic Grade Record Card. The manner in which the test scores of a pupil on a number of standardized tests can be shown graphically by means of grade equivalents is illustrated in Figure 31. This figure is representative of a simple type of test record card. It not only reveals the achievement of the pupil on the tests, but it also reveals at a glance the variation of ability in the different tests, and the accuracy of his classification. By changing the test scores into equivalent grade levels, a simple uniform graphic record of each pupil may be made. The following illustration will make this clear.

A pupil, Robert K., age eleven years and nine months, presents himself for entrance into a certain school system. An examination of his credentials shows that the work he has been doing in the school from which he has transferred does not parallel the work in this school system sufficiently to make his classification a simple matter. The administration of four or five well-standardized tests of achievement, together with one or two reliable group intelligence tests, will provide data of much value in placing this child in the new school system at a level at which he will be able to do his best work and which will be thoroughly fair to him. An intelligence test and achievement tests in reading, arithmetic, language, writing, and spelling are used, and the following scores are earned:

PUPIL · ROBERT K.	
Test	Score
Intelligence	118
Reading	38
Arithmetic	25
Language	21
Writing (Quality)	57
Spelling	61

Table XLV presents grade norms for the various tests which Robert K. has taken. The norms are hypothetical throughout, and are presented merely for illustrative purposes.

TABLE XLV
GRADE NORMS FOR INTELLIGENCE AND ACHIEVEMENT TESTS

Grade	Intelligence	Reading	Arithmetic	Language	Writing	Spelling
4	81		13		50	32
5	96	23	21	10	55	47
6	109	32	28	16	59	57
7	120	40	32	21	64	66
8	129	46	35	25	70	74
9	136	50	36			

By referring the score made on any of these tests to the table of grade norms, the grade level of achievement nearest which the accomplishment of this child places him can be determined. For example, a score of 38 on the reading test places him three-fourths of the way between the norms for the sixth and seventh grades. Accordingly, this score may be thought of as 6.7 on the grade level scale. In a similar way the arithmetic test score is four-sevenths of the distance between the norms for the fifth and sixth grades, and this score may be considered as 5.6 on the grade level scale.

The graphic record card given in Figure 31 presents the record of this pupil. At the time the tests were given he was in 5A grade. On the intelligence test he made a score almost equal to a seventh grade child, in reading he scored midway between the sixth and seventh grade norms. On all but two tests, arithmetic and writing, he exceeded sixth-grade norms. Because of his superior mental ability and his excellent work in reading and language, and in spite of his somewhat lower scores in arithmetic and writing, it was thought advisable to give him an opportunity to attempt to carry 6A work. Many other similar uses for this type of graphic record may be worked out by the interested teacher.

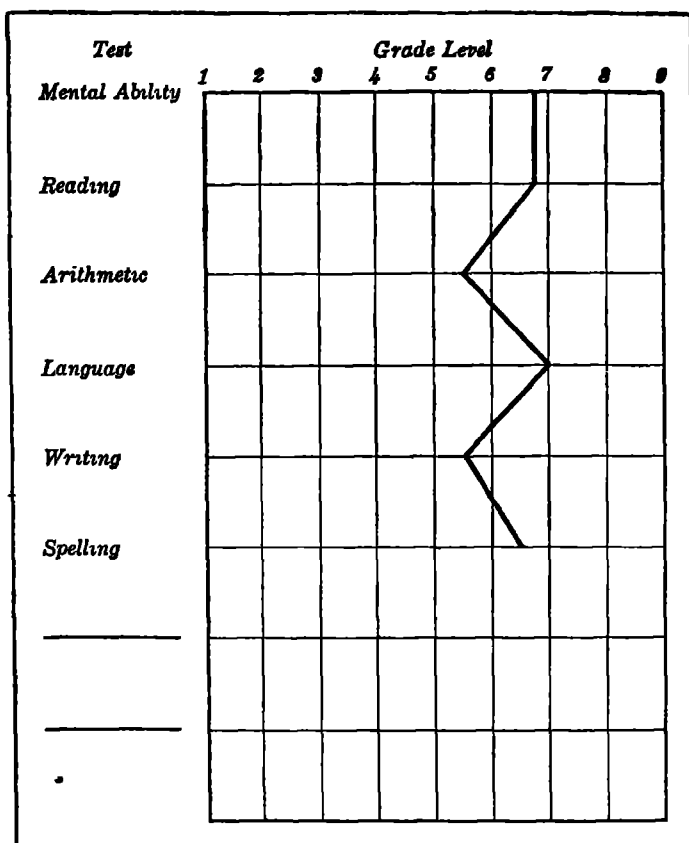


FIGURE 31 PUPIL'S GRAPHIC RECORD

Interpretation of General Achievement Test Results.

Age norms are particularly valuable where it is desirable to make use of the test results of individual pupils rather than of entire classes or grades. The illustration of age equivalents in Table XLVI is based upon data accompanying the *Stanford Achievement Test*. On the basis of these the achievement age of a pupil may be obtained without regard to his grade classification. For example, suppose a pupil of eleven years and three months makes a composite score of 63 on the advanced examination. A reference to this table immediately reveals the fact that he has made a score which

is typical of a large group of pupils of twelve years and five months. This age, 12-5, thus becomes the educational or achievement age of this pupil. Since the pupil's actual age is only eleven years and three months, this means that his achievement is one year and two months above the typical score of pupils of his age.

TABLE XLVI
AGE EQUIVALENTS OF EQUATED SCORES FOR PARTS AND
TOTAL OF STANFORD ACHIEVEMENT TEST ⁵

Equated Scores	Age Equivalents	Equated Scores	Age Equivalents	Equated Scores	Age Equivalents
79+	16-0+	59	11-7	39	9-1
78	16-0	58	11-5	38	9-0
77	15-10	57	11-3	37	8-11
76	15-7	56	11-1	36	8-9
75	15-4	55	10-11	35	8-8
74	15-0	54	10-10	34	8-7
73	14-9	53	10-8	33	8-6
72	14-6	52	10-6	32	8-5
71	14-3	51	10-5	31	8-4
70	14-0	50	10-3	30	8-3
69	13-9	49	10-2	29	8-3
68	13-6	48	10-0	28	8-2
67	13-3	47	9-11	27	8-1
66	13-1	46	9-9	26	8-0
65	12-10	45	9-8	25	7-11
64	12-8	44	9-7	24	7-10
63	12-5	43	9-6	23	7-9
62	12-2	42	9-5	22	7-9
61	12-0	41	9-3	21	7-8
60	11-10	40	9-2	20	7-7

Interpretation of Diagnostic Test Results. An illustration of the manner in which results from a diagnostic test can be interpreted by the use of norms is given for the *Compass Diagnostic Tests in Arithmetic*, Test 7, *Multiplication of Fractions and Mixed Numbers*. Table XLVII presents

⁵ Adapted from *Class Record and Class Analysis Chart Stanford Achievement Test*, Intermediate and Advanced Batteries, p. 2 World Book Co., Yonkers-on-Hudson, N. Y., 1940.

the scores made on this test by pupil R. H., age eleven years and three months, who is in the fifth grade. Tables XLVIII and XLIX give grade and age norms respectively for this test.

TABLE XLVII

PUPIL RECORD FROM COMPASS DIAGNOSTIC TEST No VII

Part	1	2	3	4	5	Total
Score	17	13	11	30	4	75
Age Equivalent			.			10 yrs. 9 mos
Grade Equivalent	H6	H6	H6	-5	-5	

Grade Equivalents. From the norms of Table XLVIII, it can readily be determined that the score of 17 on Part 1 represents achievement at the H6 or L7, meaning high sixth- or low seventh-grade levels. Similarly, scores of 13 and 11 on Parts 2 and 3 represent H6 or L7 achievement levels. For Parts 4 and 5, however, scores of 30 and 4 respectively are below the fifth-grade level, and therefore achievement is rated at -5 for these parts. For purposes of individual diagnosis this is the significant information. This pupil is in need of further practice on the multiplication of fractions. Possibly his lack of experience with exercises calling for the finding of errors may account for his relatively low score on that function.

Age Equivalents. The age equivalent of ten years and nine months (10-9), obtained from Table XLIX for a total score of 75 on the test, indicates that pupil R. H. is at the level on multiplication of fractions typical for pupils ten years and nine months of age. As he is eleven years and three months old, retardation of six months is shown by this fact. His retardation is apparently accounted for by his deficiencies on Parts 4 and 5 of the test.

Interpretation of English Test Results. The percentile norms given in Table L will serve as the basis for an illustration of the use of percentile norms for the high school level. The percentile norms for the *Rinsland-Beck Nat-*

TABLE XLVIII
GRADE NORMS FOR COMPASS DIAGNOSTIC TEST No. VII⁶

Part	Grades						
	H5	L6	H6	L7	H7	L8	H8
1	16	16	17	17	17	17	18
2	12	12	13	13	14	14	15
3	9	10	11	11	12	12	12
4	35	42	49	53	57	62	67
5	8	9	10	11	12	12	13
Total	80	89	100	105	112	117	125

TABLE XLIX
AGE EQUIVALENTS OF TOTAL SCORES FOR
COMPASS DIAGNOSTIC TEST No. VII⁷

Scores	Age Equivalents	
	Years	Months
66-70	10	6
71-75	10	9
76-79	11	0
80-84	11	3
85-89	11	6
90-94	11	9
95-98	12	0
99-102	12	3
103-105	12	6
106-109	12	9
110-112	13	0
113-115	13	3
116-117	13	6
118-120	13	9
121-124	14	0
125-128	14	3
129	14	6

⁶ G. M. Ruch, F. B. Knight, H. A. Greene, and J. W. Studebaker, *Manual of Directions for Compass Diagnostic Tests in Arithmetic*, Table 7, p. 50. Scott, Foresman and Co., Chicago, 1925.

⁷ *Ibid.* Table 22, p. 55.

ural Test of English Usage are furnished in detail for the total score and in brief form for the three test scores.

A high school senior has taken this test and made scores of 67 on Test I, *Mechanics*, of 72 on Test II, *Grammar*, and on 45 on Test III, *Rhetoric*. His total score is therefore 184. By reference to the upper portion of Table L, it can be determined that his total score places him at about the 88th percentile among high school seniors, and that, consequently, only about 12 percent of twelfth-grade pupils score higher than he did on this test. This conclusion is reached by reading across the table in the row for a score of 180 to the column numbered "4." In the cell so located, representing a score of 184, the 87.6 value indicates a percentile score closer to 88 than to any other number.

Indications of very general diagnostic significance can be obtained by the use of the lower half of the table. It appears from a comparison of his scores on Tests I, II, and III with the quartiles in the table that he scored at about the 75th percentile or upper quartile on Test I, considerably above that level on Test II, but not much above the 50th percentile or median on Test III.

Interpretation of Intelligence Test Results. Table LI gives mental age norms for the *Pmtner General Ability Tests* which will be used to illustrate the method of obtaining the mental age (MA) and intelligence quotient (IQ) from an intelligence test score.

If a pupil who is twelve years seven months of age makes a standard score of 151 on the *Pmtner General Ability Test*, reference to Table LI will show that he has a mental age of twelve years and one month (12-1). It is apparent immediately that he is somewhat below average in intelligence. If it is desired that his relative brightness be determined, his intelligence quotient can be computed from the facts at hand. The formula

$$IQ = 100 \frac{MA}{CA}$$

becomes for his case,

$$IQ = 100 \frac{12-1}{12-7} = 100 \frac{145 \text{ (months)}}{151 \text{ (months)}} = 96.$$

TABLE L
 PERCENTILE NORMS FOR HIGH SCHOOL SENIORS,
 RINSLAND-BECK ENGLISH USAGE TEST^a
 For Testing at the End of the School Year
 (FOR TESTS I, II, AND III COMBINED)

Score	0	1	2	3	4	5	6	7	8	9
200	99 0	99 2	99 2	99 2	99 2	99 5	99 5	99 7	99 7	99.9
190	94 6	95 3	96 3	96 8	97 5	97 5	97 8	98 3	98 5	99 0
180	82 5	84 2	84 7	86 6	87 6	88 5	90 2	92 4	93 2	94 1
170	65 5	69 4	71 8	73 7	75.7	76 9	77 6	78 6	79 8	81.3
160	49.2	50 7	52 1	53 6	55 5	57 5	59.7	60 6	61 8	62 8
150	34 9	36 1	38 3	39 8	41 5	43 2	43 6	45 1	47 0	47 5
140	22 5	23 7	24 7	26 6	27 4	29 3	30 3	31 3	31.7	32 2
130	13.3	13 5	14 5	15 0	15 7	16 0	17.7	19 1	20.6	21 6
120	7.0	7 7	8 0	8 7	9 2	9 7	10 9	11 4	11 8	12 3
110	3 6	3 8	3 8	4 3	4 6	4 6	4 8	5 5	5 8	6 5
100	1 9	2 4	2 6	2 6	2 6	2 9	2 9	2 9	3 3.	3 6
90	2	2	2	2	2	4	7	9	14	1.6
80	.0	0	0	.0	0	0	0	2	2	2

(FOR THE SEPARATE TESTS)

	Test I	Test II	Test III	Total Test
Upper Quartile	66 6	64 4	48 7	174 6
Median	60 6	59 8	43 5	161 5
Lower Quartile	53 2	51 6	37 4	143 1

^a Henry D Rinsland and Roland L Beck, *Teacher's Manual Rinsland-Beck Natural Test of English Usage*, Table V, p 10 Public School Publishing Co., Bloomington, Illinois

TABLE LI

MENTAL AGE VALUES CORRESPONDING TO STANDARD SCORES,
PINTNER GENERAL ABILITY TESTS^a

Mdn. Stand. Score	Mental Age	Mdn. Stand. Score	Mental Age	Mdn. Stand. Score	Mental Age	Mdn. Stand. Score	Mental Age
100	6-11	125	9-1	150	12-0	175	15-7
101	7-0	126	9-2	151	12-1	176	15-9
102	7-1	127	9-3	152	12-2	177	16-0
103	7-2	128	9-5	153	12-3	178	16-2
104	7-3	129	9-6	154	12-5	179	16-4
105	7-4	130	9-7	155	12-6	180	16-6
106	7-5	131	9-9	156	12-8	181	16-9
107	7-6	132	9-10	157	12-9	182	17-0
108	7-7	133	9-11	158	12-11	183	17-2
109	7-8	134	10-0	159	13-0	184	17-4
110	7-9	135	10-1	160	13-2	185	17-7
111	7-10	136	10-2	161	13-3	186	17-10
112	7-11	137	10-4	162	13-5	187	18-0
113	8-0	138	10-5	163	13-7	188	18-3
114	8-1	139	10-7	164	13-9	189	18-6
115	8-2	140	10-8	165	13-10	190	18-8
116	8-3	141	10-9	166	14-0	191	18-11
117	8-4	142	10-10	167	14-3	192	19-2
118	8-5	143	11-0	168	14-4	193	19-5
119	8-6	144	11-2	169	14-6	194	19-8
120	8-7	145	11-3	170	14-8	195	19-11
121	8-8	146	11-5	171	14-10	196	20-2
122	8-9	147	11-7	172	15-0	197	20-5
123	8-10	148	11-8	173	15-2	198	20-8
124	9-0	149	11-10	174	15-5	199	21-0

^a *Directions for Administering and Scoring Pintner General Ability Tests, Intermediate and Advanced*, Table I, p. 5. World Book Co., Yonkers-on-Hudson, N. Y., 1938.

PROBLEMS IN INTERPRETING TEST SCORES

PROBLEM 24

Finding Grade Equivalents from Test Scores

Use the table of grade levels adapted from the norms for the *Thorndike-McCall Reading Scales*, Table XLIV, and complete the work begun in this list of pupil scores. The first two are filled in for illustration.

Pupil	Score	Grade Level	Pupil	Score	Grade Level
1	4	2 4	7	8	
2	26	7 7	8	18	
3	6		9	28	
4	30		10	12	
5	15		11	22	
6	1		12	14	

PROBLEM 25

Finding Age Equivalents from Test Scores

Use the table of age equivalents for the *Stanford Achievement Test*, Table XLVI, and complete the work begun in this list of pupil scores. The first two are filled in for illustration.

Pupil	Composite Scores	Age Equivalent	Pupil	Composite Scores	Age Equivalent
		Yr. Mo.			Yr. Mo.
1	25	7 11	9	79	
2	34	8 7	10	66	
3	20		11	30	
4	65		12	44	
5	51		13	37	
6	41		14	55	
7	57		15	70	
8	36		16	48	

PROBLEM 26

Finding Percentiles from Test Scores

Use the upper part of the table of percentile scores for the *Rinsland-Beck Natural Test of English Usage*, Table L, and complete the work begun in this list of pupil scores. The first two are filled in for illustration.

Pupil	Total Score	Percentile	Pupil	Total Score	Percentile
1	120	7	6	199	
2	169	63	7	142	
3	133		8	98	
4	129		9	171	
5	155		10	207	

PROBLEM 27

Finding Intelligence Quotients from Test Scores and Pupil Ages

Complete the work begun in the accompanying list of pupils and scores. Use the accompanying table of age norms for the *Haggerty Intelligence Examination, Delta 1*, to determine age equivalents. The first two quotients have been correctly computed for illustration.

REVISED AGE NORMS FOR
INTELLIGENCE EXAMINATION, DELTA 1¹⁰

Year	Month											
	0	1	2	3	4	5	6	7	8	9	10	11
6	30	32	34	36	38	40	42	44	46	48	50	52
7	54	55	56	57	57	58	59	59	60	61	62	63
8	64	65	66	66	67	68	68	69	70	70	71	71
9	72	73	74	74	75	76	76	77	78	78	79	79
10	80						.					

¹⁰ M. E. Haggerty, *Manual of Directions Haggerty Intelligence Examination, Delta 1*, p. 8 World Book Co., Yonkers-on-Hudson, N. Y., 1929.

Pupil	Chronological Age		Point Score	Age Equivalent		Intelligence Quotient
	Yrs.	Mos.		Yrs.	Mos.	
1	7	0	54	7	0	100
2	7	2	59	7	6	105
3	6	6	63			
4	8	5	58			
5	7	6	55			
6	7	0	36			
7	7	4	55			
8	8	1	42			

SELECTED REFERENCES

- Broom, M. E., *Educational Measurements in the Elementary School*, Chapter IV. New York McGraw-Hill Book Co., Inc., 1939.
- Conrad, Herbert S., "Norms" *Encyclopedia of Educational Research*, pp. 773-80. New York The Macmillan Co., 1941.
- Courtis, S. A., "The Interpretation of Scores in Tests and Examinations." *Journal of Educational Research*, 31 637-49, May 1938.
- Crawford, J. R., *Age and Progress Factors in Test Norms*. Research Studies in Educational Measurements I, University of Iowa Studies in Education, Vol IX, No 4. Iowa City University of Iowa, 1934.
- Cureton, E. E., "The Accomplishment Quotient Technic." *Journal of Experimental Education*, 5 315-26, March 1937.
- Franzen, Raymond, *The Accomplishment Quotient Technique*. Contributions to Education, No. 125. New York Bureau of Publications, Teachers College, Columbia University, 1922.
- Greene, Harry A., *Work-Book in Educational Measurements*. New York Longmans, Green and Co., 1937.
- Judd, Charles H., *Educational Psychology*, Chapter 28. Boston: Houghton Mifflin Co., 1939.
- Lang, Albert R., *Modern Methods in Written Examinations*, Chapter XI. Boston Houghton Mifflin Co., 1930.
- Lee, J. Murray, *A Guide to Measurement in Secondary Schools*, Chapter XII. New York D Appleton-Century Co., Inc., 1936.
- Lincoln, Edward A., and Workman, Linwood L., *Testing and the Uses of Test Results*, Chapters V-VI. New York The Macmillan Co., 1935.
- Nelson, M. J., *Tests and Measurements in Elementary Education*, Chapter XIV. New York The Cordon Co., 1939.
- Orleans, Jacob S., and Sealy, Glenn A., *Objective Tests*, Chapter IX. Yonkers-on-Hudson, N. Y.: World Book Co., 1928.

- Ross, C. C., *Measurement in Today's Schools*, Chapter X. New York Prentice-Hall, Inc., 1941.
- Ruch, G. M., "Minimum Essentials in Reporting Data on Standard Tests" *Journal of Educational Research*, 12 349-58; December 1925.
- Seder, Margaret, *Introduction to Testing and the Use of Test Results* New York Educational Records Bureau, July 1940.
- Smith, Henry L., and Wright, Wendell W., *Tests and Measurements*, Chapter IV. New York Silver, Burdett and Co., 1928.
- Symonds, Percival M., *Measurement in Secondary Education*, Chapters XIII, XV. New York The Macmillan Co., 1928.
- Tiegs, Ernest W., *Tests and Measurements for Teachers*, Chapter XII. Boston. Houghton Mifflin Co., 1931.

CHAPTER XXIV

USING THE RESULTS OF TESTING

This chapter, in a sense, represents the objective of all of the preceding discussion of the problems of classroom measurement and evaluation. It is concerned with the problems of what should follow the testing program. In general, it assumes the point of view that any time or money spent in the construction or purchase of testing equipment is wasted unless the results revealed by the tests are returned to the classroom. Pupil and teacher time spent in any kind of testing situation is time wasted unless some tangible aid comes back to the classroom which will improve the conditions under which the pupils learn and the teacher teaches. The laws of compensation hold in measurement and evaluation in the classroom as in other fields.

I. PUPIL ADJUSTMENT THROUGH MEASUREMENT

Education as Guidance. One of the very important outcomes of testing is the establishment of a more adequate basis for the guidance of the pupil while he is in school and in the selection of his later life activities.

The guidance function of education assumes a prominent place in the more modern conceptions of the aims of education. The teacher is the guide who accompanies the youth through his educational journey and sees to it that he is given an opportunity to encounter as many as possible of the types of life situations which he is likely to meet later as a more mature youth and as an adult. The guide is responsible for giving assistance when it is needed. It is assumed that the guide has been over the route before and knows the most interesting and important side-trips and the most economical short-cuts. The guide may be assumed also to know just which of the experiences sampled out of life really are the important ones. This places a very severe responsibility upon the teacher, and it may be that the load is too great. Possibly it is too much to expect that many teachers will be able to serve as efficient guides when they are often so inex-

perienced in the ways of life that they themselves need guidance.

Inexperienced and sometimes poorly trained teachers and inadequate courses of study combine to make it important that reasonably reliable and valid supplementary devices be used to supply the essential guidance information. Ideally the best possible guide a pupil could have is a wise, sympathetic, cultured, widely-traveled teacher with a sound philosophical background and a workable psychology of learning. The absence of such a person in a school system makes it necessary to seek the required guidance in other quarters. Even the best judgment of the most sympathetic and learned teachers is subjective and is frequently wrong. Accuracy, objectivity, and validity in the data upon which the pupil's guidance is to be based are apparently essential also. Accordingly, it appears that the results of objective measurements have an important place in the educational program.

In a general way the guidance functions of education are represented in two forms: (1) the directive and (2) the corrective. The directive function of the school program is to provide the proper opportunities for the learning of the important activities which life is likely to call upon a person to perform. The corrective aspects of education involve the more or less definite admission that something is wrong in the previous educational experience and that this difficulty can be remedied only through a further modification of the educational treatment one receives. This may be based upon an unfortunate assumption, but it is true that very few pupils pass through our school systems without at some time revealing a very definite need for corrective treatment of one form or another. The use of educational test data as the basis for administration of the corrective program in education constitutes the major consideration of this section of the chapter.

Learning is fundamentally an individual and personal matter. Each person learns for himself through his own efforts, and each person probably learns most effectively in his own peculiar way. It is the responsibility of the school and of the teacher to discover as far as possible these individual peculiarities and limitations of pupils. This responsibility is likely to be more keenly felt when the results of

the pupil's efforts are unsatisfactory. When learning is accomplished with little friction and pain, little attention is given to it. Far more attention is given to the analysis and diagnosis of difficulties and causes of pupil weaknesses than to the more positive causes of success.

Disciplinary Uses of Tests. The experienced teacher, accustomed to handling pupil adjustments in the classroom, realizes that most teacher-pupil misunderstandings arise from inadequate information on the part of one or the other. Probably by far the larger proportion of the problems of conduct in the classroom arise through the failure of the teacher to realize the actual ability of the pupil and to utilize it adequately. The active, ambitious boy of superior capacity is almost certain to get into trouble unless his energy is directed into constructive channels. Too often the instructional material to which he is exposed is so uninteresting as to bore him. Many times he already knows as much about the subject as his teacher. In such cases it is small wonder that he gets into disciplinary difficulties.

Educational tests are not necessarily proposed as the cure-all for every type of disciplinary case, but it is quite certain that through their use the teacher will find much help in securing a better understanding of such cases. Difficulties arising out of superior capacity inadequately utilized or stimulated may be anticipated through the analysis of intelligence test and achievement test results. Undue dissipation of time and energy, through giving too much attention to extra-school activities, may be revealed in terms of achievement test results. Failure to achieve because of limited capacity may also be accounted for by such methods. Classroom teachers in all subject-matter fields, through the use of properly selected educational tests, may not only improve their understanding of the pupils but may also secure a most effective basis for the stimulation of the individual pupils to accomplish at their own most effective levels.

Classification and Placement of Pupils. Instructional efficiency within the classroom depends to a considerable degree upon the accuracy with which the material to be taught is adjusted to the ability of the learner. With the present rigid organization of curricular material in accordance with

arbitrarily defined grade lines, it is quite important that pupils be classified within the grade in accordance with certain definite principles. Pupils will learn most effectively when placed with other pupils who have approximately the same initial abilities and who are able to learn at about the same rate. This means that pupils of approximately the same capacity and achievement levels should be grouped together for instructional purposes. For the accomplishment of this purpose intelligence tests and general achievement tests naturally afford the basic data.

For classification purposes the use of the mental-age and achievement-age scores is recommended, since it results in placing together persons who have reached approximately the same levels of mental development. The ordinary basis of classification, the chronological age, disregards completely the actual power of the individual to learn. Grouping of pupils within narrow chronological-age limits actually results in a wide variation in mental ability. This is serious, since it is the mental ability not the chronological age of the individual which principally determines his learning rate. Grouping pupils within narrow mental-age limits also results in wide variations in life age, but this is not so serious. Other factors which should be considered in the classification and placement of pupils are the judgments of the teachers, and such objective evidence as may be gained concerning the physiological, social, and moral normality of the individual.

Sectioning of Pupils A second factor in the placement of the pupil in the group for learning purposes is to place together individuals who are able to progress at approximately the same rate. While the mental age (plus data on the educational level of the pupil, and his physiological, moral, and social development) is very useful in locating the pupil as to general grade classification, it does not afford an adequate basis for sectioning the pupils within the grade. Every teacher knows that the instructional problems within the class are greatly increased when the range of ability is wide. Under these circumstances special efforts must be made to make the instruction simple enough for the least able pupil. To accomplish this for the slow pupil means that the more able ones are likely to be bored. Instruc-

tion adjusted, on the other hand, to stimulate the better pupils goes completely over the heads of the slow ones. Accordingly, it appears that in all cases of classes large enough in size to warrant more than one section for instructional purposes, the solution lies in placing together individuals who are at approximately the same level of mental maturity or are possessed of practically the same aptitudes in the subject.

An illustration of the manner in which results from a special aptitude test may be used for this purpose is given in the accompanying tabulation. The data in Table LII represent the scores on the *Iowa Algebra Aptitude Test* from an entire entering ninth-grade class in a small high school. According to the percentile norms for this test, pupils who score as low as the twentieth percentile are almost certain to fail or encounter great difficulty in algebra. The six individuals scoring 31 points or less may, therefore, very properly be diverted into other fields in which they have a more reasonable chance of successful accomplishment. This leaves one hundred five pupils to be divided into three sections. A practical division of these pupils would be to place in a superior or more rapidly moving section the thirty-three pupils whose scores are 62 points or more on the aptitude test, and to place the thirty-five pupils scoring from 35 to 49 points in a more slowly moving section. The middle group of thirty-seven pupils may be expected to make progress at

TABLE LII
SCORES FROM A NINTH-
GRADE CLASS ON THE
IOWA ALGEBRA APTITUDE TEST

Score	Pupils
86 - 88	1
83 - 85	1
80 - 82	0
77 - 79	3
74 - 76	4
71 - 73	3
68 - 70	6
65 - 67	6
62 - 64	9 33
59 - 61	10
56 - 58	Average 8
53 - 55	Section 10
50 - 52	9 37
47 - 49	7
44 - 46	9
41 - 43	10
38 - 40	6
35 - 37	3 35
32 - 34	0
29 - 31	Divert 2
26 - 28	from the 1
23 - 25	course 2
20 - 22	0
17 - 19	1 6
Total ..	111

about the normal speed for this subject. Such a plan should result in a low percentage of pupil failures in the course, due to the elimination of the very poorly equipped individuals at the outset and also to the very specific information relative to pupil weaknesses which the teacher will have in advance.

Evaluation of Teaching Method. Another very important use of educational test results lies in the evaluation of methods of instruction. As a matter of fact, it is only since the development of educational tests to their present state of refinement that any significant evaluation of teaching methods or materials has been feasible. Such a procedure naturally involves a rather carefully devised experimental situation in which all of the variables with the exception of the one under observation are controlled as far as possible. A preliminary measurement designed to equate the experimental groups or to reveal initial abilities must be taken. Following this the experimental factor is applied. A final measure of a type similar in every basic respect to the initial measurement is then taken. Differences in the initial and final results for the two groups under comparison are taken as indications of the operation of the experimental factor.

II. MEASUREMENT OF PROGRESS AND IMPROVEMENT BY STANDARDIZED TESTS

Norms as Essentials in Measurement of Progress. One of the features which distinguishes standardized tests from other objective measures of achievement is the fact that they are uniformly accompanied by norms. Norms not only are tangible evidences of refinement in the development of the test, but they afford the basis for the practical interpretation of the meaning of the test results. The importance of the existence of suitable norms and the necessity for unusual care in their derivation and interpretation have been given special emphasis in an earlier chapter in this book. The purpose here is to make it quite clear that they are essential in all cases in which the tests are to be used for the measurement of progress. Non-standardized test results and the subjective ratings assigned by teachers to essay-type examina-

tions provide no basis for the measurement of progress, since there is no point of beginning from which to express progress.

In subject-matter fields in which the units of instruction follow each other in rather systematic and logical order, some work has been done recently in the development of narrow-unit tests which are standardized at the end of the period of instruction on the unit. In certain other subjects which do not lend themselves to this type of analysis into narrow units, there is a growing tendency to provide norms which are adjusted to fit practically any period in the school course when the tests may be administered. Norms of this type make it possible to reveal progress and improvement in a very definite manner.

Limitations of Standard Tests as Measure of Improvement. The accuracy of measurement of improvement, growth, or progress is essentially conditioned by all of the limitations inherent in the tests themselves. To these may be added certain limiting qualities which appear in the norms for the tests. No measurement can be more meaningful than the validity the instrument itself provides. Improper emphases, inadequate samplings, ambiguous exercises, subjective scoring methods, improper working time limits, and doubtless many other factors limit the accuracy of all educational measurement. Measurement of improvement requires the use of test norms. Accordingly, any inherent inadequacies, such as the lack of suitable groupings, a failure to select representative populations, the use of insufficient population samplings, or errors in computation of the norms themselves, will also operate to limit the accuracy of measures of improvement as well as any other interpretations which may be attempted.

III. MEASUREMENT BY INFORMAL OBJECTIVE TESTS

Informal Objective Tests as Measures of Class Progress. The inadequacy of the standardized educational test to meet all types of measurement uses has been consistently emphasized in this volume. There are certain functions which the standard test simply cannot perform effectively in

the classroom. This limitation is inherent in its construction. The very fact that the standard test is designed to cover the common elements in the course of study makes it practically certain that the test will be unsuited for use in any class in which a different course of study is taught. Furthermore, the standard test is usually prepared by an individual other than the instructor of the class in which it is to be used. Thus, there is little or no assurance that the emphasis which the teacher has given to the subject matter will be reflected in the test content. Usually the standard test covers a much wider area of the subject matter than the instructor desires to cover in his periodical measurements. While these limitations of the standard test concern themselves mainly with matters of the validity of the test content, they are sufficient to invalidate its use for many purposes.

The problems of constructing and using informal objective tests have been given considerable attention in this volume. However, it may not be entirely useless to remind the teacher once more that without doubt by far the larger portion of time and effort he gives to the problems of measurement will involve the use of instruments other than standardized tests and scales. Whether or not the devices used are objective or subjective in character will largely determine the significance of the results.

Subjective Bases of Teachers' Marks. The case against the subjective evaluation of classroom accomplishment has been developed in current educational literature to the point that apparently little remains to be said. The evidence all points clearly to the fact that estimates based on personal impressions and unsupported by objective criteria may be expected to be quite unreliable. A large part of this unreliability is accounted for by the absence of tangible units of measurement and of standards of expectancy on the part of the observer. The best evidence shows that a reasonably high validity may be found in many examinations utilizing the discussion-type question. The difficulty seems to be to secure an accurate estimate of the quality of the discussion. Something of the difficulty teachers have in the assignment of subjective marks to pupils' papers is demonstrated by the

data in Table LIII. This table summarizes the marks assigned by a group of 557 classroom teachers and teachers in training to a geography paper written by a sixth-grade boy.

This paper, in the opinion of the teacher of the class, was a passing paper, passing in this case being understood to be 75 or better. It is not the fact that the average of the marks assigned by these teachers is below passing which is the startling feature of this array. It is the extremely wide and consistent disagreement of the individuals as to the quality of the paper. Similar data selected from more exact fields, such as arithmetic, show distressingly similar results.

It is quite probable that the preparation and use of carefully evaluated answer keys and the administration of special periods of training in the marking of such examinations might eliminate some of the subjectivity.

Refinements in methods of evaluating pupil achievement which do not definitely place an objective unit of measurement in the hands of teachers are incomplete at best. Since it is apparent that by far the larger part of the evaluation of achievement is almost certain to be on the basis of teachers' judgments, it is important that really effective devices be used. For the time being at least, it appears that the best solution of this problem involves the more extensive use of objective examination techniques and the wider popularization of simple statistical methods of turning test scores into meaningful teachers' marks.

Functions of the Teacher's Mark. Since the teacher's mark is now and almost certainly will continue to be the most

TABLE LIII
MARKS ASSIGNED TO A SIXTH-
GRADE GEOGRAPHY PAPER
BY 557 INDIVIDUALS

Marks Assigned	No of Teachers
89 - 91	1
86 - 88	4
83 - 85	5
80 - 82	17
77 - 79	14
74 - 76	40
71 - 73	57
68 - 70	95
65 - 67	102
62 - 64	75
59 - 61	66
56 - 58	40
53 - 55	24
50 - 52	9
47 - 49	4
44 - 46	2
41 - 43	1
38 - 40	1
Total	557

frequently recorded measure of pupil accomplishment, it is very important that the classroom teacher have a definite notion of the functions of such marks. Teachers' marks function in a four-fold way: (1) they provide the basis for the school's record of the child's educational history, (2) they furnish the teacher with a record of the pupil's achievement and progress, (3) they reveal to the pupil the school's evaluation of his effort and accomplishment, and (4) they furnish to the parent reasonably accurate information concerning the pupil's achievement. For the school administrator, marks afford the common basis for determination of promotions, scholastic honors, and school classification. For the teacher, marks provide a working basis for group distinctions in assignments, work requirements, extra-curricular activities, etc. For the pupil, marks should give accurate information concerning the amount and quality of work done. For the parent, the marking system should supply accurate information on pupil achievement which should indicate relative success or failure in unmistakable terms. Obviously the realization of these four functions of the marking system places it under a very severe burden. The real severity of this burden is better appreciated when one recalls the implications of the experimental evidence on the reliability of teachers' marks, and then, in the face of these disturbing facts, realizes the seriousness with which these marks are taken by the pupil, by the parents, and even by the school itself.

Objectifying the Marking System. A critical examination of the marking system and the marks assigned by teachers makes it very apparent that some radical improvements in these phases of educational measurement are needed. As the result of an extensive survey of the problem, and a study of the recommendations of educators who have studied the problems of the marking system, the following program for eliminating many of the unsatisfactory features of the present methods of assigning marks is submitted. In brief, the plan is as follows:

1. *Discard the practice of marking pupils in percentages.* Three reasons are advanced for this decision (a) The percentage scale has for its only fixed points 0 and 100. The

former means just no ability while the latter means perfect mastery. Yet the complete scale is rarely used in practice (b) The establishment of the limits of the scale fixes the intermediate values. Accordingly, the difference between marks of 75 and 76 should be the same as the difference between marks of 97 and 98. Common observation reveals the absurdity of this assumption. (c) The use of the percentage scale presupposes that the teacher is able to distinguish as many as 101 minute differences in accomplishment. Experimental evidence¹ reveals that teachers are able to distinguish not more than five to seven levels of ability. The use of a finer scale implies an exactness of discrimination on the part of teachers which does not exist. (d) The use of an arbitrarily selected percentage as a passing mark, as is very common practice, results in throwing the marks into a badly skewed distribution with too large a proportion of the marks piled up at or near the passing mark.

2. *Each mark assigned to a pupil should be a symbol designed to indicate his power to do.* This symbol should be defined in exactly worded statements, understood alike by teachers, administrators, and pupils.

The following definition of letter marks² is cited to illustrate the type of statements that may be used to describe the qualities of an individual deserving classification in each of the letter steps in the five-point scale. •

MARK

DEFINITION

- A
1. Consistently does more than is required
 2. Has wide vocabulary at his command
 3. Is always alert, takes active part in discussions.
 4. Has unusual dependability in taking assignments.
 5. Is prompt, neat, and thorough in all work and unusually free from teachers' correction.
 6. Knows how to select books, tools, materials, and is a rapid worker.
 7. Has initiative and originality in attacking problems.
 8. Has ability to associate and re-think the problem and can adapt himself to new and changing situations.

¹ G M Ruch, *The Objective or New-Type Examination*, pp 370-74 Scott, Foresman and Co, Chicago, 1929

² E K Hillbrand, "A High School Marking System" *School and Society*, 21 142-43, January 31, 1925

9. Has enthusiasm for and interest in his work.
10. Has ability to apply ideas gained in study to **everyday** life.

- B**
1. Frequently does more than is required
 2. Has good vocabulary and speaks with conviction.
 3. Unusually alive in meeting the situation at hand.
 4. Careful in complying with assignment.
 5. Eager attack on new problems, profits from criticism.
 6. Prompt, neat, thorough, and usually accurate in all work.
 7. Has ability to apply general principles of the course.
- C**
1. Does what is required
 2. Possesses a moderate vocabulary.
 3. Willing to apply himself during class hour.
 4. Does daily preparation with comparative freedom from carelessness.
 5. Attentive to assignments.
 6. Has ability and willingness to comply with instructions and a cheerful response to correction.
 7. Reasonably thorough and prompt in all work.
 8. Has average neatness and accuracy in all work.
 9. Has ability to retain collectively the general principles of the course.
- D**
1. Usually does what is required.
 2. Attendance often irregular.
 3. Tools and equipment sometimes lacking.
 4. Frequently "misunderstands" assignment.
 5. Willing but slow in complying with instructions and corrections.
 6. Careless in preparation of assignments.
 7. Lacking in thoroughness and sometimes tardy with work.
 8. Careless in presentation of work.
- F**
1. Usually does a little less than is required.
 2. Listless and inattentive in class.
 3. Tools and equipment for work often lacking.
 4. Always tardy with work.
 5. Seldom knows anything outside the lesson.
 6. Retains only fragments of the general principles of the course
 7. Lacking in qualities of the first three groups to the extent that he cannot or will not do the work.

3. *Each teacher should give objective examinations or quizzes frequently throughout the term, and the scores from these tests should afford the major basis for his marks.* Prior to the assignment of marks for a school period or semester, composite period or semester scores should be determined for the pupils and these composite scores should

then be transformed into marks on a five-point letter scale by the use of the standard deviation technique in the case of large sections or classes (thirty or more). In the case of small classes (less than thirty), this may be accomplished somewhat more simply by dividing the distribution of scores into five groups and assigning the designated marks to previously determined percentages of the class. The letter marks used and the typical percentages of the class assigned each under these conditions are as follows:

Letter Marks	A	B	C	D	F
Percentage of Class	5-8	20-25	34-50	20-25	5-8

The essential steps in the assignment of marks by the standard deviation method are outlined in Chapter XXIII. The actual solution of a problem utilizing this method in the assignment of marks to objective test scores from a class of forty-five pupils is shown in Table XLIII, page 562.

4. *Require teachers to prepare in advance for each six-weeks' period carefully worded statements of the objectives of each subject for that period.* Unless this is done, no one can determine whether or not the pupils are being tested over the things on which they should be tested. This statement of objectives should be the criterion by which the validity of the objective tests is determined.

5. *Work prepared for daily assignments should be treated as a requirement of the course, but marks assigned should be determined by numerous brief objective quizzes or tests over the work assigned.*

6. *Notebook and laboratory work should be treated as a requirement of the course, and credit should be deducted or withheld for work which is unsatisfactory or incomplete.* However, the marks assigned should be determined by frequent objective tests over the work rather than on the basis of the notebook or laboratory work, which may or may not be the pupil's own products.

TABLE LIV
SUGGESTED POINT VALUES CORRESPONDING
TO LETTER MARKS

Mark	Points
A+	16
A	15
A-	14
B+	12
B	11
B-	10
C+	8
C	7
C-	6
D+	4
D	3
D-	2
F	0

7. Assign marks on accomplishment or performance rather than on indefinite subjective factors such as effort, attitude, ability, etc.

8. Final marks summarizing all of the quiz and test scores for the course can be obtained quite readily by assigning point values to each letter mark, computing the actual average for each pupil, and then assigning the final class marks on the basis of these averages. This is a very simple way of assigning final marks for fairly large groups and in courses in which a relatively large number of objective measures are to be summarized in the final mark. It also permits the weighting of certain period and final tests in accordance with the teacher's judgment of their importance.

Table LIV, showing point values corresponding to specific letter marks, may be useful to the teacher. Weightings are suggested for plus and minus values of the letter marks as one means of softening some of the shock from the arbitrari-

ness of letter marks assigned on the basis of the normal curve. Pupils whose test scores fall just below the point where a superior mark is given sometimes feel that this is a distinct element of unfairness in the system. Assigning the plus and minus values to their quiz scores serves to take care of this problem quite adequately.

SELECTED REFERENCES

- Bangs, C. W., and Greene, H. A., *Teachers Marks and the Marking System* University of Iowa Extension Bulletin, No. 244. Iowa, City University of Iowa, May 15, 1930.
- Brueckner, Leo J., and Melby, Ernest O., *Diagnostic and Remedial Teaching*, Chapter I. Boston Houghton Mifflin Co., 1931.
- Odell, C. W., *Educational Measurement in High School*, Chapter XIX, New York The Century Co., 1930.
- Orleans, Jacob S., *Measurement in Education*, Part II. New York. Thomas Nelson and Sons, 1937.
- Rinsland, Henry D., *Constructing Tests and Grading in Elementary and High School Subjects*, Chapter III. New York McGraw-Hill Book Co., Inc., 1938.
- Ruch, G. M., *The Objective or New-Type Examination*, Chapter III. Chicago Scott, Foresman and Co., 1929.
- Ruch, G. M., and Stoddard, George D., *Tests and Measurement in High School Instruction*, Chapters II-III. Yonkers-on-Hudson, N. Y. World Book Co., 1927.
- Rugg, Harold O., "Teachers' Marks and Marking Systems" *Educational Administration and Supervision*, 1 117-42, February, 1925.
- Wilson, Guy M., and Hoke, Kremer J., *How to Measure* (Revised and Enlarged Edition), Chapter XXVI. New York The Macmillan Co., 1928.
- Woody, Clifford, and Sangren, Paul V., *Administration of the Testing Program*. Yonkers-on-Hudson, N. Y.. World Book Co., 1932.

CHAPTER XXV

TESTS AND THE CLASSROOM TEACHER

This chapter presents a brief summary of a few of the more outstanding problems involved in the use of tests and the interpretation and application of test results by the classroom teacher as presented in the foregoing chapters of this volume.

I. THE NEED FOR MEASUREMENT

Recognition of the Need for Tests in the Classroom. Teachers long have measured the results of their teaching efforts. However, only relatively recently has any large degree of accuracy been injected into their methods of measurement. For many years the teacher's estimate was accepted as the sole measure of a pupil's ability or accomplishment. Studies of the reliability of such methods gradually cast a doubt on their accuracy. Accordingly, interested teachers and research workers began a search for more dependable measures. This movement was fostered by the so-called "survey" movement among educators which appeared at about the same time. Possibly the survey movement itself was a product of the same spirit of unrest and dissatisfaction with educational methods which brought into being the measurement movement. The survey movement left in its wake a perfectly logical result—the establishment of many centers of interest in the possibilities of the more exact evaluation of the results of educational practices. These later appeared as bureaus of educational measurements and research, many of which are still functioning. Very distinctive service has been rendered by these agencies through their work in the construction, standardization, and critical evaluation of educational measuring instruments.

The testing movement has passed through the first stages of its development. Twenty years ago it was necessary to popularize the idea. Now the advantages of standardized and informal objective tests are recognized by most educators and by many laymen. Moreover, the tests themselves are

being subjected by these same persons to careful analysis and refinement. The most enthusiastic students of educational tests are their own severest critics. The shortcomings of tests are coming to be realized. They are not mysterious instruments for the confusion of the uninitiated, but are useful devices for assisting the progressive educator to improve the conditions under which teachers teach and children learn. If they aid in the accomplishment of this, they are justified.

Meaning of Educational Tests. The meaning of educational tests to the classroom teacher may be made clear most readily by considering the characteristics which distinguish them from other types of measuring instruments in education. In the first place, the educational test of standardized or semi-standardized form is more limited in its usefulness than the informal objective examination. That is to say, it usually confines itself to the general aspects of the subject which can be covered by all classes while the typical examination or objective test prepared by the classroom teacher covers a specific selection of the subject matter dealt with in the teacher's own class. In the second place, the exercises comprising standardized educational tests are commonly constructed and arranged in accordance with certain statistical and educational principles designed to produce more accurate measuring instruments. In the third place, the more useful standardized educational tests have been used by a large sampling of school children under controlled conditions. From this use of the tests the norms which give meaning to test results and permit the interpretation of test scores are derived. In the fourth place, the more carefully constructed and valuable educational tests yield results which point the teacher's way to the application of specific remedial methods where needed.

Educational tests are characterized by other features such as validity, reliability, objectivity, etc. Validity refers to the truth of the picture of the ability or achievement revealed by the test. Reliability refers to the consistency with which the test reveals this picture. Objectivity refers to the extent to which the test results are affected by the personal judgment of the user. There are other important characteristics

of educational tests, but these represent the major ones on which their meaning depends.

II. OBJECTIVE NON-STANDARDIZED TESTS

Teacher-Made Measures of Achievement. The emphasis on the use of the standardized test in most discussions of measurement problems often leads to the mistaken idea on the part of the student that these more formal types of tests are the most important measures of achievement. In most subject-matter fields this is distinctly not the case. The use of some form of testing procedure for instructional purposes probably constitutes nine-tenths of the teacher's measurement activity in the classroom. Accordingly, much more attention should be given to the improvement of the teacher's informal measures of achievement.

Using Informal Objective Tests in the Classroom. The so-called new-type examination has increased in popularity with great rapidity in the past decade, although it is well recognized that there are certain aspects of educational accomplishment which it does not measure adequately. The advantages claimed for the informal objective examination are as follows:

1. Extensive sampling
2. Objectivity of scoring.
3. Economy of time
4. Elimination of bluffing.

Some of the more important criticisms of objective examinations which have been suggested are as follows:

1. Neglect of training in organization and expression of thought.
2. Overemphasis upon factual knowledge.
3. Encouragement of guessing
4. Difficulty of preparation
5. Considerable cost

The successful construction of objective examinations calls for the application of many of the same principles of test construction that are involved in the development of standardized tests. In many respects informal objective and standardized tests are quite similar in nature and in use.

While results from non-standardized tests within certain limits yield information which may aid the teacher in making instructional adjustments within the class, a major practical value lies in their use as the basis for class marks. Standardized test scores are not at all suitable for this purpose, but informal tests of the objective type afford the ideal basis for the objectification of the marking system.

III. STANDARDIZED EDUCATIONAL TESTS

Uses of Standardized Educational Tests. Educational tests, because of their definiteness and objectivity, reveal to the teacher the status of the achievement of his class. They point out individual differences in capacity and achievement. If standardized, they set up specific goals of achievement for the teacher. They reveal the results of special types of emphasis, or of special methods of instruction. They open to the administrator and the teacher hitherto untouched sources of information useful in giving the pupil proper educational and vocational guidance. In their modern conception, they reveal to the teacher the specific weakness of individual pupils so definitely that he is in a position to apply effective instructional and corrective methods. Tests themselves have little or no power to bring about changes in pupil achievement as a mere result of their use. Their chief service is their power to reveal pupil strengths or weaknesses. The correction of weaknesses is another aspect of the supervisory problem.

Using Standardized Tests in the Classroom. Teachers themselves must assume a larger share of the responsibility for the use of educational tests in the classroom, and for the interpretation and application of the results after the tests have been given. Only by so doing does the teacher receive an adequate return from the use of tests. If this responsibility is to be wisely assumed, the teacher must have an appreciation of the possibilities and the weaknesses of tests. He must be trained in their use and the interpretation of their results. He must be willing to exchange a certain amount of personal effort for the information concerning his teaching problems which the tests can furnish him.

Training in the use of tests comes as a result of their use. Opportunity for this training may be afforded through the preparation and use of objective informal tests as substitutes for the traditional examination, or it may be provided by undertaking a study of some supervisory problems of importance to the teacher in which standardized tests are used.

Main Uses of Tests. Three main types of uses of educational tests are noted, each resulting in different points of view regarding the teacher's responsibility. Tests of a detailed diagnostic type designed to give the teacher precise information concerning the abilities and limitations of his pupils are instructional in their function. The responsibility for the use of such material should be the teacher's. Tests designed to be used more particularly for survey or supervisory purposes should probably be administered by persons other than the teacher. The use of tests for administrative purposes such as for pupil classification, gradation, or sectioning, may well be the joint responsibility of the teacher and the administrator. In other words, the function the tests are intended to perform determines where the responsibility for their use and interpretation lies.

Selection of the Test to Use. The criteria for tests—validity, reliability, adequacy, objectivity, administrability, scorability, comparability, economy, and utility—afford the teacher a tangible basis for the selection of the test to use for a certain purpose. In a general way, however, the teacher should depend upon the advice of persons who have made a special study of the tests, rather than to attempt to apply these criteria personally. Information on these important aspects of tests is furnished with most of the better and more recently developed tests. If affirmative answers to each of the following questions are available concerning a specific test, the teacher may feel reasonably safe in selecting it for use.

1. Does this particular test measure the abilities, habits, skills, attitudes, and information which I wish to measure?
2. How much time does it take to give the test? Is it long enough to give a reliable and consistent measure?
3. Is it easily and accurately scored?
4. Has it been widely used elsewhere?

5. Does it furnish accurate and extensive norms for comparison and interpretation?
6. Is the interpretation of the scores simple and clear?
7. Do the results point the way to a remedial program?
8. Is the test economical in terms of time and money cost per unit of reliable information furnished by it?

Administration of the Test. One of the distinctive features of standardized tests is that they must be given under conditions closely approximating those under which they were standardized, if the results are to be meaningful. Accordingly the teacher should follow the directions furnished with such tests.

The attitude and the personality of the examiner are also important in the administration of a test. The whole purpose of the test is defeated if an unnatural response is obtained. In the giving of tests in the beginning elementary grades, the greatest care must be exercised to secure the cheerful confidence of the pupils.

Scoring of the Test Papers. Much of the value arising from the use of tests in the classroom is lost if the teacher himself does not score the papers. In the first place, the act of scoring the papers forces a closer inspection of their content than would otherwise be the case. Such careful scrutiny often reveals hitherto unsuspected features in the test. In the second place, the teacher is brought into much closer contact with the pupil's special abilities and limitations by the act of personally scoring his test paper. For the majority of the better tests, the task of scoring the papers is greatly simplified and objectified by the use of answer keys, scoring stencils, and in many cases mechanical devices.

IV. INTERPRETATION OF TEST RESULTS

Summarizing and Interpreting the Results of Testing. Skill in summarizing and interpreting test results is dependent upon the mastery of the following statistical techniques:

1. A knowledge of why and how to classify and tabulate data
2. A knowledge of how to find the common measures of central tendency.
3. A knowledge of how to express the variability of data.

4. A knowledge of how to determine the relationship between two or more groups of data
5. A knowledge of how to derive and use norms and derived scores for purposes of comparison and interpretation of test results.
6. A knowledge of how to treat data for simple graphic presentation

In addition to the use the classroom teacher may make of these skills in the proper utilization of test scores, there is the application which may be made of them in the study of current educational literature. Reports of progress in education are filled with statistical terms and techniques. The teacher can scarcely hope to keep abreast of the times in his profession if he is unable to read current educational literature understandingly.

V. PRACTICAL ASPECTS OF CLASSROOM MEASUREMENT

Diagnostic Testing and Remedial Teaching. The analysis, identification, and measurement of the abilities which underlie and condition educational achievement unquestionably constitute the high point of the use of tests in constructive supervision. Forming the background of practically all possibilities of learning is that curiously interwoven maze of inherited traits, tendencies, and predispositions known as mental ability. Naturally enough, instruments designed to sample into this field constitute an important unit of the teacher's diagnostic equipment.

Intelligence. The acceptance of the definition of intelligence as the ability of the individual to learn makes it relatively easy to set up devices for its measurement and interpretation. Intelligence tests are incapable of securing a direct measure of capacity unaffected by experience and training. They do not measure the actual process of learning, or the quality of the learning equipment directly, but provide the basis for inferences as to the equipment from the amount of learning which has taken place under certain environmental conditions. The value of the intelligence test lies in the opportunity it affords for making this inference on a reasonably objective basis. Thus the intelligence test, carefully used and critically interpreted, constitutes a most effective and useful instrument for classroom diagnosis. Not only do intelligence test scores provide valuable evidence

as to basic or general limitations and superiorities, but they frequently offer most helpful hints as to the existence of more highly specialized abilities or disabilities. In the last analysis, predictive tests which render such important service in certain types of educational and vocational guidance are specialized tests of intelligence.

Personality. In the sense that an individual's personality is evidenced in all of his behavior, this aspect of classroom measurement is all inclusive. In a somewhat narrower sense, personality has to do with such forms of behavior as attitudes, interests, and emotional adjustment, all of which are important considerations in the classroom. Personality inventories and scales are doubtless in their early stages of development, but they afford evidence of types not realized from intelligence or achievement tests which teachers should find valuable in the guidance and adjustment of their pupils.

Tool Subjects. From the discussion of the diagnostic and remedial possibilities in the field of arithmetic, it is apparent that here is a subject in which some rather definite analytical and constructive work is being done. It is now possible to diagnose arithmetic disabilities with considerable accuracy in the fields of whole numbers, fractions, decimal fractions, percentage, denominate numbers, mensuration, interest and business forms, and problem solving. Effective remedial and corrective materials are also available in whole numbers, common fractions, decimal fractions, and problem solving. The near future will undoubtedly see the rapid development of more effective instructional and corrective exercises in these as well as in other aspects of arithmetic.

The subject of reading has undergone many interesting and significant developments in recent years, largely because of a more adequate realization of the social importance of certain types of reading skills. The fact that the main bulk of all reading done under present social conditions is of the work-study type has been recognized in the rapid shift in instructional emphasis from the oral to the silent reading field. This does not mean that there is not an important place for the development of oral reading skills, but that they will be developed in their own place and for certain purposes. Obviously these purposes are not in conflict with

those underlying the development of silent reading skills. The recognition of these specific functions of the various types of reading has made it possible for much progress to be made in the analysis of reading disabilities. Reading tests capable of furnishing results accurate enough for individual diagnosis are now available for such reading skills as word meaning, sentence meaning, paragraph meaning, sentence organization, paragraph organization, use of an index, alphabetizing, and rate of reading. Considerable progress has also been made in the construction of drill materials intended to increase skill in these phases of reading ability.

The diagnostic and remedial program in language is decidedly less clean-cut than in certain other subjects. This is unquestionably due to the present lack of definiteness in the identification of the fundamental language skills. As yet, relatively little has been done in the objective identification of oral language disabilities, and but little more has been done in the development of corrective instruction in this field. Written language in certain forms may be measured with reasonable accuracy, but the causes lying at the foundation of difficulty in this field are still indifferently identified. Certain more specific language skills, such as capitalization, punctuation, language usage, and certain grammatical situations, can now be diagnosed with a fair degree of accuracy. However, the discouraging feature of this situation is that remedial instruction based on well-constructed drill on such specifics as punctuation, capitalization, language usage, etc., all of which, in theory at least, should contribute to success in the use of written language, fails to improve the general quality of the child's written language product. This means that one of two things is true—either our criterion of language success is inadequate, or the specific skills on which instructional emphasis is placed are not the proper ones. It must then be apparent that, while the language problem is an old one, it is still with us. Much more work needs to be done in the development of more accurate analyses, more valid and reliable diagnostic instruments, and more effective corrective and instructional materials in the field of language.

Spelling and handwriting, two of the very important mechanical elements in written language, have both been

analyzed and measured with reasonable success. The course of study content in spelling is becoming quite well known. The importance of providing the pupil with an adequate technique for learning to spell is generally recognized. The functions of testing in spelling instruction are apparent. The fact that by far the larger portion of testing in spelling is done in connection with teaching may ultimately carry over into other subjects.

Handwriting is a product of classroom and individual activity rather than the result of a pupil's response to stimuli. Products generally require evaluation by means of scales or score cards, and as a result the values assigned are frequently highly subjective. However, writing has responded to analysis sufficiently to enable the classroom teacher and supervisor to set up a rather definite corrective program in spite of the subjective elements involved in its measurement.

Content Subjects. Mental abilities and the special tool subjects lend themselves reasonably well to analysis and diagnosis. The content subjects, such as the social sciences and the more exact sciences, however, because of vagueness in the statements of their aims and purposes, are extremely difficult to measure in an analytical manner. Mastery of the tool subjects, such as reading, language, etc., enters into the measures of achievement obtainable in these fields to such a degree that many times the results are practically meaningless. However, as more critical analyses are made of curricular materials in these subjects, and when the aims and outcomes of instruction are more exactly stated, measurement in these fields is certain to be made more accurate and analytical.

Fine Arts. Changes in educational emphasis from the vocational and practical to the cultural, and increases in opportunities for the enjoyment of leisure are combining to force increased attention to the fine arts. That public interest in good music and in the applied forms of art is being rapidly awakened is indicated by the many state and sectional music contests and in the numerous community art projects now under way. The progressive classroom teacher of these subjects will wish to utilize all of the suggestions as to con-

tent, method, and outcomes which the tests in these fields make available.

Health and Physical Education. The modern emphasis upon preventive measures in health education and adaptation of physical education to individual needs is motivation for many of the modern evaluative techniques in this area. However, it is in evaluative devices of a non-test type rather than in standardized tests that this trend is most clearly evidenced. Although the teacher most concerned is the one directly in charge of activities of this nature, health is so fundamentally important that all teachers should be conversant with some of the tests and simple evaluative procedures.

General Educational Achievement. While the emphasis throughout this volume is definitely upon the measurement of the specific rather than the general aspects of school accomplishments, there is a recognizable need for the latter type of measurement. For general survey purposes, for evaluation of curricular content, and for later individual detailed diagnosis, such general achievement tests are valuable. The important point which the teacher should keep in mind, however, is that such measurement is only the beginning, not the end, of classroom testing. Throughout all of the supervisory and instructional uses of educational tests it must be remembered that vague and indefinite measurement is useless, that measurement which stops short of identifying the existence and causes of individual difficulties is futile, and that correction of defect without exact diagnosis is practically impossible.

Measurement and the Total Child. There are certain aspects of ability, accomplishment, skill, aptitude, character, and personality which unquestionably lend themselves to reasonably objective measurement. The emphasis which is given to these measurable qualities frequently gives the impression that they represent the major elements in the total understanding of the child. Such is far from the case, however, for many of the intangibles of the child's personality are almost certainly of greater importance, although in many cases they are practically impossible to measure objectively. This merely means that the teacher must be made keenly aware of the fact that something lies beyond

objective measurement. He must see that appraisals of the child's total personality are basic to effective classroom teaching. He must recognize that many (probably most) of these vital appraisals must be made on the basis of keen observation and sympathetic analysis of his pupils. Even if the teacher were gifted with unusual observational and analytical power, superior native capacity, and natural sympathy, even if he were a four-year college or normal school graduate with graduate degrees in pedagogy, psychology, sociology, psychiatry, and medicine, he could not hope to comprehend more than a few problems of the child's personality. The important point here is that, while it is impossible to know all, it is not impossible for the teacher to be cognizant of and sensitive to these problems.

After Testing, What? This question is in the back of the mind of every classroom teacher and every supervisor who has used standardized tests. Much of the early use of tests was futile, since such broad vague phases of educational achievement were tested that, even though reliable results were obtained, nothing specific could be done about the situation. Furthermore, a great deal of the early use of tests in the classroom was a matter of satisfying curiosity. Hundreds of thousands of dollars' worth of tests have been given to millions of children, taking thousands of hours of pupil and teacher time, with meagre returns. Teachers have a right to expect that something tangible will be given them in return for interruptions and for pupil time spent in testing. Pupils themselves may even have some rights in the matter. One way to insure this return is for the teachers themselves to take an active part in the program. A type of training, an attitude toward their profession, a clearer insight into the difficulties faced by their pupils, are thereby gained which may not come to them in any other way.

The results of supervisory tests given periodically for the purpose of checking the efficiency of the teacher's instruction should be revealed to the classroom teacher in terms of specific suggestions for the further improvement of the situation. Instructional and diagnostic tests used by teachers in the classroom should furnish such specific information concerning the abilities and limitations of their pupils that a

program of preventive and corrective instruction can be begun at once.

Although a great deal of attention is now being focused on the problems of developing corrective and remedial materials to use for the follow-up program after testing, only a limited amount of such material is now available. It is noticeable also that most of the better corrective material is limited to two or three subjects—arithmetic, language, and reading. Illustrations drawn from these three are naturally given considerable attention in the discussion of the problems of diagnostic testing and remedial teaching.

Conclusion. Within the past quarter century, tests and measuring devices in nearly all subject-matter fields have been developed. What the future of this movement will be no one can predict. In many ways the unusually rapid development of measurement techniques was unfortunate, for in its early stages it resulted in confusion on the part of the classroom teacher, the one who should have profited most from the program. There was a danger of over-production without sufficient refinement. There has been a tendency to flood the market with more and more test devices, with only an indifferent, and mostly inadequate, attempt to answer the question of what to do after testing. Today this danger is far less serious. It is very encouraging to observe that teachers themselves are becoming better informed in the techniques of testing and are demanding that the corrective and remedial aspects of objective supervision shall materialize in the classroom.

SELECTED REFERENCES

- Gilliland, A. R., Jordan, R. H., and Freeman, Frank S., *Educational Measurements and the Class-Room Teacher* (Revised Edition), Chapter III. New York The Century Co., 1931.
- Orleans, Jacob S., *Measurement in Education*, Chapter XIV. New York Thomas Nelson and Sons, 1937.
- Russell, Charles, *Classroom Tests*, Chapter XVII. Boston Ginn and Co., 1926.
- Russell, Charles, *Standard Tests*, Chapter XVIII. Boston Ginn and Co., 1930.
- Wilson, Guy M., and Hoke, Kremer J., *How to Measure* (Revised and Enlarged Edition), Chapter XXVI. New York. The Macmillan Co., 1928.

GLOSSARY

- ability.** The capacity or power to produce.
- accomplishment.** See *achievement*.
- accomplishment quotient (AQ).** The ratio between educational age and mental age. $AQ = 100 EA/MA$.
- accuracy.** Ratio between number of exercises correctly done and number of exercises attempted.
- achievement.** The accomplishment or production of the child in his school work.
- achievement age.** A pupil's level of accomplishment in a particular school subject or field.
- achievement quotient.** See *accomplishment quotient*.
- achievement test.** A test which measures the pupil accomplishment resulting from school instruction.
- adequacy.** The degree to which a test samples extensively or widely over the content to be tested, an important criterion of a good examination.
- adjustment.** The process of effecting a satisfactory adaptation to one's environment.
- adjustment inventory.** An instrument used to determine how satisfactorily the individual has become adapted to various phases or a certain phase of his environment.
- administrability.** The characteristics of a test which make for ease and accuracy in giving it, a criterion of a good examination.
- age equivalent.** The score derived from age norms on a standardized test, established by determining the average score made by pupils of each age.
- age norms.** Tables of values representing typical or average performance on standardized tests for pupils in different age groups.
- alternate-response item.** A type of test item to which the pupil responds by selecting one of the two possible answers, one of which is right and one of which is wrong.
- ambiguity.** The quality of a test item which makes possible more than one logical interpretation of its intent or meaning.
- analogies test.** A test of logical reasoning ability involving similarities and dissimilarities
- analysis.** Reduction or taking apart of a total performance in the process of identifying specific skills.

analytic test. A test which furnishes a basis for the analysis of skills underlying a performance by taking several different cross-sections of abilities contributing to total performance.

anecdotal method. A technique of personality evaluation in which significant behavior incidents in the life of a pupil are noted and recorded.

anecdotal record. An objective account of pupil behavior made by the teacher or some other person observing a significant event in the life of the pupil

answer sheet. The separate sheet on which the pupil records his responses for a test

appraise. See *evaluate*.

appreciation. A judgment concerning the worth of a piece of art, an event, an experience, etc.

aptitude. Ability in a certain field or area of performance.

aptitude test. A test of specific intelligence, i.e., intelligence as it operates in a certain field or area of performance, which may be used for prognostic purposes.

arithmetic mean (A.M.). The point on the scale above which and below which the deviations are equal, the most commonly used measure of central tendency, frequently called the average.

array. A collection of data, usually organized around a particular point of reference.

association methods. Certain techniques of personality evaluation, such as the free association and projective methods.

assumed mean. The mid-point of the class-interval in which it is "guessed" that the arithmetic mean will fall, in computing the mean by the short method from a frequency distribution.

attitude A state of readiness which exerts a directive, and sometimes a compulsive, influence upon an individual's behavior

attitudes scale. An instrument used in the determination of pupil opinions or beliefs on an issue or issues which may be controversial in nature

average. A generic term covering the measures of central tendency, but commonly used to designate the arithmetic mean.

basic skills. Tool skills, such as those of reading, language, and arithmetic, essential to study of the content subjects.

behavior All types of responses made by the individual, particularly those which can be observed

- best answer item.** A type of multiple-choice item to which the pupil responds by attempting to select the best answer from alternatives of which more than one may be correct.
- bluffing.** The device used by a pupil in the attempt to convince the teacher that he knows answers to questions concerning which he is uninformed or in doubt.
- C-scores.** Derived scores based upon deviation from the average in units of one-tenth of a quartile deviation.
- capacity.** The ability to learn or profit from experience; limit of potential development.
- case study.** A comprehensive approach to the evaluation of the total personality of the individual pupil.
- central tendency.** A term corresponding to average, commonly applied to the arithmetic mean, median, and mid-measure.
- "chance-half" coefficient.** An estimate of test reliability useful when only one form of a test is available
- chronological age (CA).** Life age, the number of years since birth.
- class analysis chart.** A device for the graphical representation of class performance and individual pupil performance on the various parts of certain achievement tests.
- classification.** The process of assigning a pupil to the grade or unit of a school for which his abilities and training best fit him.
- class-interval (c i.).** One of the divisions of a frequency distribution, the column of a frequency distribution in which the limits of the tabulation units are shown.
- classroom test.** A test made by the teacher or within a school system for use in specific classes.
- clues.** Characteristics of test items which frequently aid the pupil in determining the correct answers.
- coefficient of alienation (k).** An index of the degree to which two variables are unrelated.
- coefficient of correlation (r).** A measure of relationship which ranges in value from + 1.0 through zero to - 1.0; refers here mainly to Pearson product-moment coefficient.
- comparability.** The characteristic of a test which enables the user to obtain from different administrations of the test results which have equivalent meanings; a criterion of a good examination.

comparable measures. Scores or values which are expressed in terms of the same unit and with respect to the same point of origin.

completion exercise. A type of test exercise to which the pupil responds by filling the blanks of a paragraph with the words, numbers, phrases, etc., which he believes will correctly complete the meaning.

composite score. A single value used to express the results obtained from the use of several different measures

comprehension score. A score indicating the degree of a pupil's understanding of an exercise or of material read

constant errors. Types of deviations from complete accuracy which result from the tendency of some scorers to give high marks and of other scorers to give low marks consistently.

content subjects. Fields in which mastery consists mainly in the acquisition of informations and attitudes, as the social sciences and sciences.

converted scores. Derived scores based upon deviation from the arithmetic mean in units of one-tenth of a standard deviation.

correction. An adjustment used in computing the arithmetic mean, standard deviation, and correlation coefficient by the short method from a frequency distribution or correlation chart.

correction for chance. A practice followed in scoring some types of objective tests to take account of guessing.

corrective teaching. Steps taken to remedy observed defects or difficulties in pupil learning.

correlation. The degree of relationship existing between two or more sets of measures

correlation chart A two-way or double-entry table which shows the relationship existing between pairs of measures for the same individuals or items.

correlation coefficient. See *coefficient of correlation*.

cramming. The process of attempting to learn a great deal in a short time by intensive study.

criterion. A standard by which a test or other product is judged or evaluated.

cumulative frequency. The sum of all the scores in a frequency distribution up to any given point.

cumulative pupil record. A comprehensive, cumulative record of pupil background, ability, achievement, behavior, etc., of wide usefulness in guidance.

curricular validity. Evidence of test validity shown by similar coverage of test content and curriculum content.

cursive writing. Handwriting with the letters joined.

decile. One of the ten equal parts into which a distribution of scores is divided.

derived score. A value having comparable meaning for the results from various tests and obtained by changing the form of a raw score in an established and consistent manner.

deviation. The amount by which a score or other measure differs from the central tendency of the group of scores in which it is included.

diagnosis Exact identification and location of specific strengths or weaknesses in performance.

diagnostic test. A test used to locate the nature, and if possible the causes, of disability in performance.

difficulty. The characteristic in a test item which results in a large percentage of incorrect responses.

directed observation. A technique of personality study involving observation of certain specific types of behavior in the pupil.

discriminative power. The quality of a test item which results in adequate distinctions in percentages of correct answers by pupils of varying ability levels.

dispersion. See *variability*.

double-entry table. See *correlation chart*.

drill. Repetition designed to improve skill or to make learning's permanent.

duplicate forms. See *equivalent forms*.

economy. The relatively low cost of a test; a criterion of a good examination.

educational age (EA). A pupil's level of accomplishment in a number of school subjects.

educational index. A measure of educational attainment, used in determining the difference between a pupil's achievement and his ability to achieve.

educational quotient (EQ). The ratio between educational age and chronological age. $EQ = 100 \text{ EA/CA}$

educational test. A test or scale which measures the results or effects of instruction and learning.

emotional adjustment inventory. See *adjustment inventory*.

equated scores. Derived scores which are comparable from test to test of a certain battery.

equivalent forms. Duplicate or equal forms of a standardized test which yield closely similar scores.

error of grouping. A variable error introduced by the practice of combining in class-intervals scores or measures which are unlike.

essay examination. A test to which the pupil ordinarily responds with written discussion of issues raised in several broad questions.

evaluate. To test, measure, and appraise the "whole" child by the use of tests and a wide variety of non-test techniques and devices.

examination. See *test*.

exercise. A unit of a test governed by a specific set of directions.

expectancy. The standard of future achievement held reasonable for the individual pupil.

expressive language arts. Language and grammar, handwriting, and spelling, as used here.

extensive sampling. See *adequacy*.

extrapolation. The process of locating a point beyond two or more known points in accordance with the conditions operating in the given case.

factor analysis. A method widely used in the study of the nature of mental and other abilities.

faculty theory. The theory that intelligence consists of a large number of relatively independent and largely correlated and specialized abilities, such as memory, imagination, etc.

feeble-minded. The term used to designate persons of very inferior intelligence, having IQs below 70.

fine arts. Music and art, as used here.

first quartile (Q_1). The point on a scale of values below which 25 percent of the cases fall; the 25th percentile.

"footrule" coefficient. An estimate of test reliability useful when only one form of a test is available.

fore exercise. A preliminary or practice exercise for the purpose of giving the pupil experience with the particular test situation.

form. One of the two or more arrangements of closely similar or equivalent standardized test exercises which in itself constitutes a testing unit.

free association method. A technique of personality study involving the evaluation of responses given by the pupil to certain word stimuli.

frequency (f). The number of measures in a given class-interval of a frequency distribution; the column of a frequency distribution into which the scores are tabulated.

frequency curve. A graphic representation of a distribution of measures

frequency distribution. The table in which scores or other measures are classified.

fulcrum. The axis upon which a lever is supported and rotated.

G-scores. Practically synonymous with grade scores or grade equivalents.

general ability. Closely similar to general intelligence; ability to learn.

general achievement tests. Educational tests covering several fields of subject matter and ordinarily adapted for use in several grades.

general intelligence test. A test of general mental ability.

genius. A person of very superior intelligence having an IQ of 140 or above.

gradation. See *classification*.

grade. The administrative division of the school which indicates the level of advancement of the pupil.

grade equivalent. The score derived from grade norms on a standardized test; established by determining the average score made by pupils in each grade.

grade norms. Tables of values representing typical or average performance on standardized tests for pupils in different grades.

group test. A test which can be administered to a number of pupils at the same time.

grouping. The process of classifying and tabulating data into class-intervals or steps.

guessed average. See *assumed mean*.

guidance. Aid to pupils in more wisely purposing, planning, executing, and evaluating the activities which receive their attention.

half-sum. A term used in connection with the calculation of the median; $N/2$.

"halo effect." The tendency of a teacher to be influenced in rating pupil performance by impressions previously acquired handedness. The predisposition to use one hand rather than the other in manual reactions.

idiot. A feeble-minded person having an IQ below 25.

imbecile. A feeble-minded person having an IQ from 25 to 49

index of brightness (IB). A measure of brightness somewhat similar to the intelligence quotient in meaning.

index of studiousness. The difference between a pupil's rank in his class on intelligence and on achievement.

individual differences. The observed or measured variation of individuals in ability, progress, achievement, etc.

individual test. A test which can be administered to only one pupil at a time.

informal objective test. A teacher-made objective test

instructional test. A test used directly in connection with the teaching of a unit of material.

integral limits. The lower and upper whole-number limits of a class-interval in a grouped frequency distribution

intelligence. Ability to adapt oneself to changing conditions, ability to learn.

intelligence quotient (IQ). The ratio between mental age and chronological age. $IQ = 100 MA/CA$.

intelligence test. A test which measures ability to learn or to profit from experience.

intensive sampling. A narrow and inadequate selection of test items which results in a test of too little scope or range.

interest. A mental set which urges a person in a certain direction.

interests inventory. An instrument used in the determination of pupil interests in various fields or areas of performance

interpolation. The process of locating an intermediate point between two known points in accordance with the conditions operating in the given case.

interval. See *class-interval*.

interview. A personal conference technique frequently used in diagnosis and in the evaluation of attitudes.

inventory test. A test used as a preliminary check on the degree of mastery existing prior to instruction

item count. A method used to determine whether test items properly discriminate between pupils of various ability levels.

logical validity. See *psychological validity*.

machine-scored test. A test which can be scored by the use of an electrical or mechanical scoring machine.

manuscript writing. A free-hand style of lettering in which the letters are not connected as in common script writing.

mark. The teacher's numerical or letter evaluation of pupil achievement in a course or area of performance.

matching exercise. A type of test exercise to which the pupil responds by attempting to pair the related items in two or more columns of related facts.

mean. See *arithmetic mean*.

measure. To test by means of standardized and teacher-made instruments mainly in the fields of achievement and intelligence. Also a test score or other numerical rating.

median (Mdn.). The point on the scale below which half of the measures in a frequency distribution fall, a widely used measure of central tendency.

mental ability. Ability to learn, nearly synonymous to intelligence.

mental age (MA). The intelligence or mental ability of a child expressed in terms of the chronological age of which his mental ability is typical.

mental index. A measure of ability to learn, used in determining the difference between a pupil's achievement and his ability to achieve.

mental test. A test of intelligence or personality, as distinguished from an educational test

metronoscope. A device for exposing strips of reading material for reading drill.

mid-measure. The middle measure of a series of values arranged in order of magnitude, a counting measure of central tendency similar to the median.

mid-point. The exact middle of a class-interval or step in a frequency distribution.

moron. A feeble-minded person having an IQ from 50 to 69.

multiple-choice item. A type of test item to which the pupil responds by attempting to select the correct response from the several alternatives given.

multiple-response item. A type of test item differing from the multiple-choice form mainly in that the pupil responds by indicating all correct answers, of which there may be one, two, or more.

new-type examination. See *informal objective test*.

normal. Typical in progress, growth, development, or distribution.

normal curve. The graphic representation of a large number of cases in the selection of which chance was operative.

norms. The median or average performances on standardized tests of pupils of different ages or grade placement, as determined by the testing of large numbers of pupils.

objective test. A test for which the scoring procedure eliminates subjective opinion and judgment

objectivity. The characteristic of a test which eliminates subjective opinion or judgment in the process of scoring it, an important criterion of a good examination.

objectivity coefficient. A correlation coefficient used in determining the objectivity of a test.

observation methods. Certain techniques of personality study, such as directed observation and the anecdotal method.

ophthalmograph. A binocular camera used in measuring eye movements during reading.

oral examination. A test administered and answered orally.

percentile. One of the one hundred equal parts into which a distribution of scores is divided.

percentile curve. A graphical representation of the percentile points of a distribution.

percentile norms. Tables of values representing percentile ranks of scores on standardized tests for certain subjects or certain grades.

percentile rank. Position assigned to a score in an array for which the scores are divided into one hundred equal divisions in descending order.

performance. The accomplishment, achievement, or behavior of the pupil.

performance test. A test to which the pupil responds by motor or manual rather than by verbal behavior.

personal constant (PC). A measure of brightness obtained by the use of Henris growth units for both the mental age and the chronological age in the formula $PC = 100 MA/CA$.

personal reports. The types of responses given by the pupil on certain types of scales, inventories, etc., used mainly in the study of personality.

personality. An individual's total behavior in social situations.

personality inventory. An instrument which measures such intangible aspects of behavior as attitudes, interests, adjustment, etc.

personality quotient (PQ). A quotient sometimes used in the measurement of total personality.

point score. See *raw score*.

power test. A test which measures the difficulty of the task the pupil is just able to perform in terms of how far he can go through a test in which the items consistently increase in difficulty.

practice effect. The influence which previous experience with a test has on a later encounter with the same or a similar test.

practice test. See *fore exercise*.

preventive teaching. Steps taken at the time of initial instruction to guard against the later appearance of defects or difficulties in pupil learning.

primary mental abilities. The seven types of abilities of which Thurstone suggests that intelligence is constituted.

product scale. A series of items of graded difficulty, e.g., spelling words, from which tests can be constructed.

profile chart. A device used for graphical representation of scores made by the pupil on the various parts of certain achievement, intelligence, and personality tests.

prognostic test. A test used to predict future success in specific subjects or fields.

progress record. A device similar to a profile chart on which pupil progress from year to year can be shown graphically for certain achievement tests.

projective method. A technique of personality study involving the observation of how a child plays with certain toys and materials.

prophecy formula. The Spearman-Brown formula used in estimating test reliability by the "chance-half" method.

psychological examination. See *intelligence test*

psychological validity. Evidence of test validity resulting from a logical dissection of a total learning process.

quality scale. A series of standard graded samples with which the production of the pupil is compared in evaluating performance in such areas as handwriting and composition.

quartile. One of the four equal parts into which a distribution of scores is divided.

quartile deviation (Q). One of the common measures of variability or dispersion, half of the distance between Q_3 and Q_1 .

quotient. A ratio designed to reveal in a single numerical index the relative position of the pupil on two related variables.

range (R). The distance from the lowest to the highest score in a series of scores.

rate score. A score expressing a pupil's rate of work.

rate test. A test which measures speed of performance on tasks of uniform difficulty in terms of the number of exercises the pupil completes in a specified time or the time he requires to complete a specified task.

rating scale. An instrument used by a teacher or other person in the evaluation of pupil personality or achievement.

raw score. The quantitative result obtained directly from the scoring of a test or scale.

readiness test. A test which measures the ability of the pupil to undertake a new type of specific learning.

real limits. The actual or true lower and upper limits of a class-interval in a frequency distribution.

recall test. A type of test to which the pupil responds by writing words, numbers, phrases, etc., to complete the meaning, and which places responsibility for the answers upon the pupil. See *simple recall item* and *completion item*.

receptive language arts. Reading and study methods, as used here.

recognition test. A type of test to which the pupil responds by indicating the truth or falsity of statements, selecting the correct or best answer from among several given, or indicating the proper pairing of related items, and which places responsibility on the pupil only for selecting the correct responses. See *alternate-response item*, *multiple-choice item*, and *matching exercise*.

relative rank. Position assigned to a score in an array for which the scores are arranged in descending order.

reliability. The degree to which a test measures what it does measure; consistency of measurement; a major criterion of a good examination.

reliability coefficient. The correlation coefficient obtained between scores made by the same pupils on two equivalent forms of a test.

remedial. Having as a purpose the correction of observed difficulties and weaknesses in performance.

remediation. See *corrective teaching*.

retesting coefficient. An estimate of test reliability which can be obtained when only one form of a test is available.

sampling. The process of selecting a limited number of cases or items which will be representative of the large group from which they are chosen.

scale. An instrument used by the scorer in evaluating pupil performance or by the test-maker in constructing a test. Also the continuum from the lowest to the highest score in a frequency distribution.

scaled scores. Derived scores based upon deviation from the arithmetic mean in units of one-tenth of a standard deviation for a group established in a certain manner.

scaled test. A test in which the items are arranged in an order of increasing difficulty.

sciences. Such subjects as elementary science, hygiene, nature study, and general science, as used here.

scorability. The characteristics of a test which make for ease and simplicity in scoring it, a criterion of a good examination.

score. A quantitative description of performance.

self-marking test. A test which does not require the use of scoring keys or machines in the scoring process.

semi-interquartile range. See *quartile deviation*.

sigma (σ). See *standard deviation*.

simple recall item. A type of test item to which the pupil responds by writing the word, number, phrase, etc., which he believes will correctly complete a statement or answer a question.

social studies. Such content subjects as history, civics and government, and geography, as used here.

social utility. A point of view basic to the selection of curricular materials which holds that subject matter should contribute definitely to child and adult needs.

specific determiners. Characteristics of test items, such as certain words, item length, etc., which seem to determine in part the nature of the correct response to true-false items.

speed test. See *rate test*.

standard deviation (S.D.). The most widely useful measure of variability or dispersion.

- standard scores.** Derived scores based upon deviation from the arithmetic mean in terms of the standard deviation.
- standardization.** The process of constructing a test and establishing norms for it.
- standardized test.** A test for which the exercises have been carefully selected and evaluated and which is accompanied by norms.
- standards.** Levels of performance agreed upon by experts or established by local school officers as goals of pupil attainment.
- statistical validity.** Evidence of test validity shown by correlational relationships or other statistical procedures.
- step.** See *class-interval*.
- subjectivity.** The degree to which measurement results are influenced by personal opinions or judgment.
- sub-total.** A term used in connection with the calculation of the median.
- survey test.** A test which measures general achievement in certain subjects or fields.
- synthesis.** The process of combining underlying and somewhat isolated skills so that they form an effective unit.

T-scores. Derived scores based upon deviation from the arithmetic mean in units of one-tenth of a standard deviation.

tabulation. The process of grouping and classifying data for purposes of condensation and ease of interpretation. Also the distribution into which data are classified.

tachistoscope. A device for exposing strips of reading material for reading drill.

talent. See *aptitude*.

teacher-made tests. Tests constructed by the teacher, such as the essay and informal objective tests.

teacher's mark. See *mark*.

technique. A procedure or method.

telebinocular. A type of stereoscope adjustable for various distances.

test. In the general sense, any instrument used in the measurement of any educational or mental ability, in a specific sense, an instrument used by the pupil and ordinarily involving the use of paper and pencil. Also to measure by the use of tests.

test battery. A group of several tests covering a number of different subjects and intended for use in testing over wide areas.

test item. The smallest unit of a test; almost synonymous with test exercise.

test rating scales. Scales used in the evaluation of tests for specific uses.

third quartile (Q_3). The point on a scale of values below which 75 percent of the cases fall, the 75th percentile

tool subjects. Fields in which achievement consists mainly in the acquisition of skills and techniques useful in further learning, as reading, arithmetic, and spelling.

traditional examination. See *essay examination*.

true-false item. A type of alternate-response item to which the pupil responds by indicating whether a statement is true or false.

two-factor theory. The theory that intelligence consists of a general factor, many specific factors, and a number of group factors.

utility. The degree to which a test serves a definite need; an important criterion of a good examination.

validity. The degree to which a test measures what it purports to measure; the major criterion of a good examination.

validity coefficient. A correlation coefficient used in determining the validity of a test.

variability. Spread or dispersion of scores, common measures are range, standard deviation, probable error, and quartile deviation.

variable. A quality which may exist in different amounts.

variable errors. Types of deviations from complete accuracy which result from the tendency of persons to vary in their judgments from time to time.

work-type. The types of silent reading skills commonly utilized in study.

yes-no item. A type of alternate-response item to which the pupil responds by an affirmative or negative answer to a question.

z-scores. Derived scores based upon deviation from the arithmetic mean in standard deviation units.

INDEX

- Ability, definition of, 550
- Accomplishment quotient, 239-241
- Achievement quotient See *Accomplishment Quotient*
- Achievement tests, instructional uses of, 99-105
- Adequacy, of a test, 63-65
- Adjustment, and guidance, 269-271, measurement of, 257-263, personal report blanks, 261-263, rating scales, 259-261
- Adjustment inventory, meaning of, 32-33
- Administrability, of a test, 68-69
- Administration, of informal objective tests, 165-166, of intelligence tests, 223-224, of standardized tests, 114-118
- Advantages, of the essay test, 139-142, of informal objective tests, 155-158, of the oral examination, 132-133
- Age equivalents, 554-555, 574, 579, problem in finding, 579
- Age norms, 86-87
- Aldington, R., 244-245
- Algebra test, 29
- Alienation, coefficient of, 538-539
- Allen, R. D., 87, 119, 186, 279, 282, 286, 488-489, 491
- Allport, G. W., 266
- Almack, J. C., 399
- Alternate-response items, 174-177, 192-193, 412, 429, 449-450, 472
- Althouse, A. D., 179
- Ambiguity, freedom from, 78
- American Child Health Association, 466
- American Council on Education Cumulative Record Form, 274-275
- American Council on Education Psychological Examination, 212-214, 231-232
- American Handwriting Scale, 391
- Analysis, as basis of diagnosis, 290-291, class, 102-104, in language, 367-368, in oral reading, 336-339, in silent reading, 339-347
- Analytic test, meaning of, 18-22
- Analytic testings, in language, 367-368, in oral reading, 336-339
- Analytical Scales of Attainment, 310, 313
- Anastasi, A., 42, 43, 49, 51, 221
- Anderson, H. R., 419
- Anderson, R. G., 210-211, 230
- Anderson, W. N., 375, 377
- Anecdotal records, 249-250
- Answer sheets, machine-scored, 122-125
- Appel, K. E., 266
- Appraisal, meaning of, 6-7
- Aptitude test, meaning of, 28-29
- Aptitude tests, 215-216, early, 45; in music, 445-447
- Arithmetic, 304-323, bibliography, 323-325, course content, 304-309, diagnostic tests in, 313-316, general achievement tests in, 309-313, problem solving tests in, 316-318, remedial instruction in, 318-323
- Arithmetic mean, basic assumptions in computation of, 513-515, computation of, 503-509, problems in computation of, 509, summary of steps in computation of, 508
- Arithmetic skills, 306-309
- Arithmetic tests, 17, 20-21, 87, 171, 293-295, 306-318
- Army Alpha Test, 44, 210
- Army Beta Test, 44, 210, 218
- Arnold, D. L., 440
- Around the World Inventory, 253
- Art appreciation tests, 455-457
- Art education, 452-459, bibliography, 459-461, measurement of ability in, 455-459, outcomes of, 452-455
- Art judgment tests, 455-457
- Ashbaugh, E. J., 24
- Association methods, 248-249
- Assumed mean, 505
- Attitudes, measurement of, 251-254, nature of, 252, scales, 253-254, 407-408, 466
- Attitudes scale, meaning of, 31-32
- Ayer, F. C., 443
- Ayres, L. P., 41, 46, 51, 375, 389, 390
- Ayres Scale for Measuring the Handwriting of School Children, 93, 390-391, 394
- Ayres Spelling Scales, 378, 398
- B-scores, 554
- Baker, H. J., 214, 261, 286, 302
- Baker "Telling What I Do" Test, 261
- Billenger, H. L., 83, 180, 357-358, 369
- Bangs, C. W., 597
- Barker, V., 352
- Barr, A. S., 413
- Barr-Diggett Information Test in American History, 413
- Barrows, T. N., 412

- Basic Writing Vocabulary*, A, 377
 Beach, F A, 451
 Berch Standardized Music Tests, 448, 451
 Beattie, L, 323
 Beck, R L, 577
 Bedell, R C, 440
 Bell, H M, 33
 Bell Adjustment Inventory, 33, 261
 Berman, L, 245
 Betts E A, 333, 346-347, 352, 400
 Betts-Keystone Telebinocular, 346
 Betts Ready to Read Tests, 336, 346-347
 Binet, A, 43, 45, 50, 200, 229
 Binet Simon Scales, 43, 206
 Bingham, W V, 34, 221, 242
 Blanton, S, 360
 Block, V L, 284
 Bologna, University of, 38
 Bolton, F E, 136
 Bordin, E, 460
 Boston examinations, 38-40
 Bovard, J F, 476
 Boyd, W, 38
 Boynton, P L, 34, 205, 221, 242
 Brice, D K, 476
 Brainard, P P, 257
 Brainard Specific Interest Inventory, 257
 Branom, M E, 419
 Breed, F S, 376, 399
 Bregman, E O, 228
 Brightness, 226-227, index of, 231
 Brinkley, S G, 154
 Brooks, F D, 455, 459
 Broom, M E, 128, 197, 221, 242, 323, 352, 399, 419, 460, 548, 581
 Brown, A W, 413
 Brown, C M, 173
 Brown, M, 286
 Brown, M V, 286
 Brown Woody Civics Test, 413
 Brownell, W A, 72, 197, 306, 323
 Bruckner, L J, 128, 266, 302, 310, 311, 323, 324, 352, 399, 419, 476, 597
 Bruckner Diagnostic Tests in Arithmetic, 315
 Buckingham, B R, 46, 182
 Buckingham, G E, 440
 Burns S F, 460
 Buras, O K, 72, 111, 128, 221
 Burt, C L, 200
 Burton, W F, 383
 Buswell, G F, 324, 347
 Buswell-John Diagnostic Chart for Fundamental Processes in Arithmetic, 315
 C scores, 559
 Cain, M, 419
 Cildwell, O W, 38, 40, 51, 131-132, 149
 California Test of Mental Maturity, 210, 212
 California Test of Personality, 280-281
 Calkins, M W, 149
 Calvert, E F, 178
 Calvert Science Information Test, 178
 Camp, W G, 7
 Cardiovascular tests, 471
 Carman, H J, 412
 Carter, R E, 143, 150
 Carver, M, 476
 Case, A T, 253
 Case study, 285-286
 Cason, H, 149
 Casto, E R, 419
 Cattell, J M, 43
 Cattell, P, 231, 242
 Cattell, R B, 205
 Central tendency, arithmetic mean as measure of, 503-509, basic assumptions in computing measures of, 513-515, median as measure of, 510-513, mid-measure as measure of, 509-510
 Chadwick, E B, 40
 Chaille, E S, 42-43
 Chance, correction for, 166
 "Chance-half" coefficient, 62-63, 565-566, problem in computation of, 567
 Chapman, J C, 240-241
 Chassell, C F, 419
 Chassell, E B, 419
 Chave, E J, 35, 252
 Chenoweth, L B, 475
 Chinese examinations, 37
 Church, E, 460
 City testing bureaus, 108
 Civics, measurable qualities in, 404-406, types of tests in, 406-408
 Civics test, 186
 Clapp-Young Self-Marking Tests, 121
 Clark, W W, 100, 212, 281, 311, 312, 313, 353
 Class analysis, and diagnosis, 102-104, 236-237, charts, 280-283
 Class-interval, 497-501
 Class marks, assignment of, 560-563, problem in assignment of, 563
 Classroom measurement, practical aspects of, 604-608
 Classroom testing, 130-131, importance of, 151-152, need for improvement in, 131
 Classroom tests, 25-26 Also see *In formal Objective Tests and Essay Tests*
 Clues, inspiring use of, 77-78
 Coefficient, "chance-half," 62-63, 565-566, correlation, 529-540, "Footrule," 63, 566, objectivity, 67, 567, of aliena-

- Coefficient, *cont'd*
 tion, 538-539, reliability, 61-63, 565;
 retesting, 62, 565, validity, 57-59, 564
 Coefficient of correlation, computation of
 Pearson product-moment, 532-537,
 meaning of, 529-532, 537-540, practi-
 cal uses of, 564-567, problem in cal-
 culation of, 541
 Cole, R. D., 112-114
 Colestock, C., 477
 Colvin, S. S., 200
Common School Journal, 39-40
 Comparability, of a test, 69-70
 Compass Diagnostic Tests in Arithmetic,
 20-21, 293-295, 307-309, 315-318
 Completion items, 172-174, 190-192,
 410-411, 429, 474
 Composition, measurement of ability in,
 365-366
 Computational skills, measurement of
 ability in, 309-311
 Conard, E. U., 391
 Conard Manuscript Writing Standards,
 391
 Conrad, H. S., 97, 242, 581
 Constant errors, 139
 Content subjects, 607
 Converted scores, 559
 Cook, W. W., 178
 Cooper, H., 477
 Cooperative Achievement Tests, 176, 179,
 181, 184, 432
 Cooperative Achievement Tests for the
 Junior High School, 410, 430, 484-485
 Cooperative testing programs, 108-110
 Correction, 507-508, 512, for chance,
 166
 Corrective exercises See *Remedial In-*
struction
 Correlation coefficient See *Coefficient of*
Correlation
 Correlation table, tabulating test scores
 in, 532, 534
 Courtis, S. A., 38, 40, 46, 51, 131-132,
 149, 581
 Courtis-Shaw Standard Practice Tests in
 Handwriting, 196, 397, 398
 Cozens, F. W., 471, 474, 475, 476
 Cram, F. D., 89
 Crapser, A. L., 476
 Crawford, J. R., 86, 581
 Crider, B., 352
 Criteria of a good examination, adequacy,
 63-65, administrability, 68-69, com-
 parability, 69-70, economy, 70-71, ob-
 jectivity, 66-68, reliability, 61-63,
 scorability, 69, utility, 71, validity, 52-
 61
 Crow, L. D., 431
 Croxson, W. C., 441
 Cubberley, H. J., 475
 Cumulative pupil records, 273-276
 Cumulative record forms, 274-275
 Cureton, E. E., 242, 581
 Curtis, F. D., 441
 Diggett, C. J., 413
 Dilrhyple, C. O., 325
 Davis, G., 399
 Davis, I. C., 437-438, 441
 Davis, W. M., 435-436, 441
 Dein, C. D., 460
 Dearborn, W. F., 200, 242, 347
 DeGraff, M. H., 411
 Denny, E. C., 171, 185, 413, 414
 Denny-Nelson American History Test,
 171, 185, 413, 414
 Derived scores, based on average per-
 formance, 553-555, based on vari-
 ability of performance, 557-559, func-
 tion of, 552, in relation to norms,
 553, quotients as, 555-557
 Derryberry, M., 466
 Detroit Alpha Intelligence Test, 214-215
 Detroit General Aptitude Examination,
 216
 Deviations, 505-506
 Dewey, J., 316
 Diagnosis, as basis for prevention, 295-
 296, as basis for remediation, 293-295,
 class, 102-104, general vs specific,
 479-480, in arithmetic, 313-316, 317-
 318, in elementary science, 439-440,
 in handwriting, 391-395, in health
 education, 467-468, in oral language,
 360-362, in oral reading, 336-339, in
 physical education, 475, in silent read-
 ing, 339-347, in social studies, 418,
 in spelling, 379-382; meaning of, 289-
 290, nature of, 290-293, pupil, 99-
 100, 234-235
 Diagnostic Examination of Silent Read-
 ing Abilities, 143, 346
 Diagnostic test, meaning of, 18-22
 Diagnostic testing, in arithmetic, 313-
 316, 317-318, interpretation of re-
 sults from, 573-574
 Diamond, L. N., 111, 428
 Dickson, V. E., 242
 Dickerick, P. B., 7, 287
 Difficulty, of test items, 78-80
 Directed observation, 249
 Discriminative power, of test items, 80-
 81
 Dispersion See *Variability*
 Doig, D., 460
 Dolch, E. W., 352
 Double-entry table, 532-534
 Dougherty, M. L., 387, 396, 400
 Douglass, H. R., 240

- Downing, E R, 441
 Drake, R. M., 447-448
 Drake Musical Memory Test, 447-448
 Drawing scales, 455
 Drummond, A M., 360
 Dunlap, J W., 125, 256
 Dunlap Academic Preference Blank, 256-257
 Durrell, D D., 347, 352
 Durrell Analysis of Reading Difficulty, 347
 Durrell-Sullivan Reading Capacity Tests, 336
 Dvorak, A., 13, 346
 Dyer Backboard Test of Tennis Ability, 474
 Dykema, K W., 111, 399
- Economy, of a test, 70-71
 Economy Problem-Solving Exercises, 319-322
 Economy Remedial Exercises, 319-320
 Edgeworth, F Y., 133
 Edgren, H D., 476
 Educational guidance, 235-236
 Educational index, 241
 Educational measurement, periods in development of, 36
 Educational quotient, 126-127
 Educational test, meaning of, 10, 14-15, 599-600
 Educational tests, characteristics of, 2-3, essay, 25, 133-150, from 1800 to 1900, 38-41, informal objective, 25-26, 151-198, objective, 15, oral, 15, prior to 1800, 38, purpose of, 10, since 1900, 45-49, standardized, 16-24, 74-97, teacher-made, 25-26, types of, 14-26
 Eells, W C., 136
 Eginton, D P., 476
 Elementary science, 421-440, aims of, 421-422, bibliography, 440-442, diagnosis and remediation in, 439-440, informal objective tests in, 432-439, outcomes of, 422-427, standardized tests in, 427-432
 Elliott, E C., 133, 135-136, 142
 Ely, L A., 187, 411, 412, 414
 Ely-King Tests in American History, 187, 411, 412, 414
 Emotional adjustment, measurement of, 257-263
 Engelhart, M D., 72, 149, 197
 English test, interpretation of results from, 574-576
 Ephraimites, 37
 Equating test forms, 81-83
 Equivalent forms, 81-83
 Errors, constant, 139, variable, 139
 Essay questions, types of, 143-145
 Essay test, meaning of, 15, 25
 Essay tests, advantages, of, 139-142, early uses of, 37; improvement of, 143-149, limitations of, 133-139, nature of, 15, 25, steps for improvement of, 147-149
 Eurich, A C., 197, 302
 Evaluation, meaning of, 6-7
 Examinations, significance of, 5-6
 Exercises, in tabulating test scores, 502-503
 Expressive language arts, handwriting, 384-399, language, 355-373, spelling, 373-384
 Extensive sampling, 155-157
- Factor analysis, 202
 Fictual tests, in social studies, 406-407
 Faculty theory, 201-202
 Faulkner, R., 460
 Fine arts, 607-608
 Fisher, G., 40-41, 50
 Fitzgerald, J A., 352, 375
 "Footrule" coefficient, 63, 566, problem in computation of, 567
 Franseen Diagnostic Tests in Language, 367
 Franzen-Derryberry-McCall Health Awareness Test, 466
 Franzen, R., 239, 466, 581
 Free associations, 248
 Freeman, F N., 11, 34, 43, 44, 51, 200, 201, 221, 229, 231, 242, 267, 381, 386, 387, 393-394, 396, 400
 Freeman, F S., 221, 324, 352, 400, 419, 460, 491, 610
 Freeman Chart for Diagnosing Faults in Handwriting, 393, 396
 French-Cooper Volleyball Test for High School Girls, 474
 Frequency distribution, exercises in constructing, 502-503, summary of steps in constructing, 501-502, tabulation of test scores in, 494-503
 Friedberg, J., 267
 Frutchey, F P., 441
 Fryer, D., 34, 254
- G-scores, 554
 Gilton, F., 42, 49, 248
 Garrett, H L., 42, 51, 221, 526, 548
 Gates, A I., 129, 181, 287, 324, 339-342, 348, 349, 350, 352, 353, 400, 420, 465, 491
 Gates Primary Reading Tests, 340
 Gates Reading Readiness Test, 334
 Gates Silent Reading Tests, 339-342
 Gates-Strang Health Knowledge Test, 464-465
 General achievement batteries, 479-491;

- General achievement batteries, *cont'd*
 advantages and disadvantages of, 480-482, bibliography, 491
 General achievement test, interpretation of results from, 572-573
 General educational achievement, 608
 General intelligence test, meaning of, 26-27
 General intelligence tests, 205-214, group, 210-214, individual, 206-209, types of, 27-28, uses of, 234-237
 General science, objectives of, 425
 General social studies, tests in, 409-410
 Geography, measurable qualities in, 404-406, types of tests in, 406-409
 Geometry test, 176
 Gerberich, J. R., 197
 Gildersleeve, G., 450, 460
 Gildersleeve-Soper Musical Achievement Test, 450
 Gileadites, 37
 Gilliland, A. R., 221, 324, 352, 400, 419, 460, 491, 610
 Glenn-Gruenberg Instructional Tests in General Science, 428
 Goddard, H. H., 44, 206
 Good, W. R., 548
 Goodenough, F. L., 43, 221, 228
 Gradation, pupil, 100-102
 Grade equivalents, 554, 569, 574, 579, problem in finding, 579
 Grade norms, 85-86
 Grammar, measurement of ability in, 366-368
 Graphic record card, use of, 570-572
 Graves, M., 460
 Gray, C. F., 548
 Gray, C. T., 392-393
 Gray, H., 100-101
 Gray, W. H., 88
 Gray, W. S., 322, 327, 328-331, 336-339, 347, 348-349, 350, 352, 353, 394
 Gray Oral Reading Check Tests, 337-339
 Gray Score Card for Measuring Handwriting, 392-394
 Gray Standardized Oral Reading Paragraphs, 336-337
 Gray-Votaw General Achievement Tests, 101
 Greene, C. E., 124
 Greene, H. A., 20, 29, 56, 81, 89, 180, 300, 307-309, 316-317, 320, 322, 324, 344-345, 357-358, 368, 369, 400, 411, 548, 575, 581, 597
 Greenwich Astronomical Observatory, 42
 Greenwich Hospital School, 40
 Griffin, J. B., 287
 Griffiths, N. L., 217
 Grim, P. R., 419
 Grimes, J. W., 460
 Group comparisons, 104
 Group intelligence test, meaning of, 27-28
 Group intelligence tests, 210-214
 Grouped data, computation of arithmetic mean from, 504-508, computation of median from, 510-513, computation of quartile deviation from, 518-521, computation of standard deviation from, 524-528
 Grover, C. C., 334-335
 Groves, E. R., 267, 287
 Gudakunst, D. W., 476
 "Guess Who" Test, 263
 Guidance, 270-273, and adjustment, 269-271, education as, 583-585, educational, 235-236, pupil, 99, vocational, 236
 Guidance tests, 283-284
 Haggerty, L. C., 176
 Haggerty, M. E., 176, 259-260, 580
 Haggerty Reading Examination, 176
 Haggerty-Olson-Wickman Behavior Rating Schedules, 259
 Half-sum, 510-511
 "Halo-effect," 138
 Hamalainen, A. E., 419
 Hand, H. C., 283-284
 Hand-scored tests, 118-121
 Handwriting, 384-399, bibliography, 399-401, diagnosis of disability in, 391-395, measurable qualities in, 385-388, measurement of ability in, 388-391, remedial instruction in, 394-399
 Handwriting scale, 23
 Harris, A. J., 332, 333, 353
 Harrison, M. L., 353
 Harry, D. P., Jr., 183-184
 Hart, E. H., 441
 Hartshorne, H., 267
 Hawkes, H. E., 53, 73, 128, 161, 163, 174-175, 197, 400, 419, 441
 Hayes, M., 260-261
 Hayes Scale for Evaluating the School Behavior of Children, 260-261
 Health attitudes inventories, 466
 Health education, 462-468, 608, aims of, 463-464, bibliography, 476-478, measurement and evaluation in, 464-467; prevention and diagnosis in, 467-468
 Health knowledge tests, 464-465
 Heath, M. L., 477
 Heath-Rodgers Soccer Test for Elementary School Boys, 474
 Hutton, K. L., 7
 Heins, H., 230
 Henmon, V. A. C., 200
 Henri, V., 43

- Hereditary Genius*, 42
 Herndon, A., 267, 287
 Herrick, V E., 353
 Herring, J P., 206
 Hilden, A H., 230, 231
 Hildreth, G H., 128, 217, 312
 Hildreth Arithmetic Achievement Tests, 312
 Hillbrand, E K., 593-594
 Hilligas, M B., 46, 365
 Hilligas Composition Scale, 365
 Hilpert, R S., 453
 Hippocrates, 245
 History, measurable qualities in, 404-406, types of tests in, 406-408
 History tests, 171, 185, 186, 187
 Hoke, K J., 325, 354, 401, 420, 478, 597, 610
 Holzinger, K J., 221, 548
 Home economics test, 173
 Horn, E., 328-331, 350, 375, 377, 379, 383, 400
 Howe, E C., 477
 Howland, A. R., 477
 Huey, E B., 347
 Huffaker, C L., 240
 Hull, C., 34, 221
 Hunt, T., 34, 51, 221, 267, 287
 Hygiene, objectives of, 424
 Illinois Examination, 182, 239, 490
 Index, educational, 241, mental, 241, of brightness, 231, of studiousness, 241
 Individual diagnosis, 234-235
 Individual differences, 37-38, 41-42, 271
 Individual intelligence test, meaning of, 27
 Individual intelligence tests, 206-209
 Informal objective test, meaning of, 15-16, 24-26, vs standardized test, 153-155
 Informal objective tests, advantages of, 155-158, as measures of class progress, 589-590, construction and use of, 160-169, content of, 161-163, development of, 47-49, limitations of, 158-160, preparation of, 163-164, use of item files for, 167-169
 Intelligence, 604-605, distribution of, 232-233, early measurement of, 42-43, indirect measurement of, 203-204, measurement of, 202-205, nature of, 200-201, theories concerning, 201-202
 Intelligence quotient, 226-230, 576, 580, constancy of, 228-229, future of, 229-230, problem in finding, 580-581
 Intelligence test, meaning of, 10-11, 26
 Intelligence testing, from 1800 to 1900, 41-43, since 1900, 43-45
 Intelligence tests, administration and scoring of, 223-224, care in use of results from, 224-225, derived results from, 225-232, factual content of, 204-205, general, 205-214, interpretation of results from, 576-578, performance, 218-220, purpose of, 10-11, specific, 215-218, types of, 26-30
 Interests, inventories, 245-257, measurement of, 254-257, nature of, 254
 Interests inventory, meaning of, 32
 Intermediate School Auto Mechanics Test, 179
 International Test Scoring Machine, 122, 124
 Interpreting test results, 126-127
 Interview, 252, 285
 Iowa Algebra Aptitude Test, 29, 587
 Iowa Elementary Language Tests, 568
 Iowa Every-Pupil Test in General Science, 430
 Iowa Every-Pupil Tests of Basic Skills, 17, 120, 177, 293-295, 310, 368, 370-372, 484-485, 491
 Iowa General Information Test in American History, 411
 Iowa Grammar Information Test, 89
 Iowa Language Abilities Test, 83, 180, 368-369
 Iowa Placement Examinations, 215
 Iowa Revision of the Brace Test of Motor Ability, 470, 471
 Iowa Silent Reading Tests, 56, 342-345, 368, 491
 Iowa Spelling Scales, 24, 377, 378
 Irwin, M E., 441
 Jastrow, J., 43
 John, L., 324
 Johnson, G B., 471
 Jones, A J., 271, 287
 Jordan, R H., 221, 324, 352, 400, 419, 460, 491, 610
 Judd, C H., 97, 332, 581
 Jung, C G., 245
 Junior American History Test, 412
 Kandel, I L., 133, 149
 Kefauver, G N., 283-284
 Kefauver-Hand Social-Civic Guidance Test, 283-284
 Kelley, T L., 171, 173-174, 205, 242, 419, 430, 491, 548
 Kelley, V H., 56, 344-345
 Kellogg, C E., 31, 218-219
 Kellogg, M., 310, 313
 Kellogg-Morton Revised Beta Examination, 31, 218-219
 Kelly, F J., 145
 Kern, M R., 460
 King, E., 187, 411, 412, 414

- Kinney, L. B., 197
 Kinter, M., 460
 Kirby, C. V., 453-454, 460
 Kirby Grammar Test, 366
 Klapper, P., 324
 Klar, W. H., 453, 460
 Kline-Carey Drawing Scales, 455
 Knauber Test of Art Ability, 456
 Knight, E. W., 37, 38
 Knight, F. B., 20, 300, 305, 307-309, 316-317, 320, 322, 324, 575
 Knight-McClure Arithmetic Neatness Scale, 391
 Koos, L. V., 387, 400
 Kopel, D., 354
 Kraepelin, E., 49
 Kramer, E. E., 548
 Kretschmer, E., 245
 Krenz Test Scorer, 121-122
 Krey, A. C., 419
 Krug, E. A., 441
 Kuder, G. F., 63, 566
 Kuhlmann, F., 44, 206, 210-211, 230
 Kuhlmann-Anderson Intelligence Tests, 210-211, 230
 Kunitz, A., 477
 Kwalwasser, J., 443, 444, 449, 450, 451, 460
 Kwalwasser Test of Music Information and Appreciation, 449, 450
 Kwalwasser-Dykema Music Tests, 446
 Kwalwasser-Ruch Test of Musical Accomplishment, 450, 451

 Lancaster, F., 419
 Lang, A. R., 34, 38, 51, 73, 97, 128, 150, 197, 548, 581
 Lange, S. R., 419
 Language, 355-373, analysis of ability in, 367-368, analysis of skills in, 356-359, bibliography, 399-401, oral, 358-362, remedial instruction in, 369-373, social importance of, 355-356, written, 362-365
 Language arts, expressive, 355-401, receptive, 326-354
 Language tests, 83, 89, 171, 177, 180, 181, 293-295
 LaPorte, W. R., 469
 Larson, W. S., 460
 Laycock, S. R., 287
 Leamer Practice Sentences in Handwriting, 397, 398
 Learning, measuring efficiency of, 104-105
 Leary, B. E., 464
 Lee, D. M., 351
 Lee, J. M., 35, 48, 97, 128, 166, 197, 267, 302, 353, 548, 581
 Lee, R. E., 440
 Lee-Clark Reading Readiness Test, 335
 Lefever, D. W., 197
 Lessenger, W. E., 240
 Levine, A. J., 221
 Lewerenz, A. S., 456-459
 Lewerenz Test in Fundamental Abilities of Visual Arts, 456-459
 Lewis, D., 446
 Lewis English Composition Scales, 366
 Limbert, P. M., 253
 Limitations, of the essay test, 133-139, of informal objective tests, 158-160, of the oral examination, 132
 Limited sampling, 134-135
 Lincoln, A., 389
 Lincoln, E. A., 35, 128, 197, 302, 548, 581
 Lindquist, E. F., 53, 62, 63, 73, 128, 161, 163, 174-175, 197, 400, 419, 441, 548
 Linc, W., 287
 Lippincott-Chapman Classroom Products Survey Tests, 490
 Lockhart, A., 477
 Long, J. A., 176
 Lowman, C. R., 477
 Ludemann, W. W., 287
 LuPone, O. J., 433-435, 441

 McAdory Art Test, 455, 456
 McBroom, M., 328-331, 357-358
 McCall, W. A., 47-48, 51, 97, 130, 197, 264-266, 302, 466
 McCall Inter-Trust Rating Scale, 264-266
 McCallister, J. M., 420
 McCauley, C., 449
 McCauley Examination in Public School Music, 449
 McCloy, C. H., 470-471, 472-475, 477
 McGill, A., 412
 McGill Every Pupil Test of Geography, 412
 McKee, P., 353, 357-358, 400
 Machine-scored answer sheets, 122-125
 Machine-scoring devices, 122-125
 Madsen, I. N., 129, 221, 242, 324, 353, 400, 420
 Maladjustment, causes and symptoms of, 258
 Miller, J. B., 267, 441
 Mann, C. R., 53, 73, 128, 161, 163, 174-175, 197, 400, 419, 441
 Mann, H., 39-40, 50, 131-132
 Manuul, H. T., 443
 Mirks, L., 221
 Marshall, H., 241
 Martin, W. A. P., 37
 Master Achievement Tests, 171
 Matching exercises, 182-187, 194-196, 413-415, 431-412, 450, 471
 Mathematics test, 181, 183-184

- Maurer, K M, 228
 Maxfield, F N, 232
 Maxwell, G W, 311, 312
 May, M A, 267
 Mead, A R, 129
 Measurement, in education, 2, meaning of, 6-7, not new idea in education, 1
 Measuring efficiency of learning, 104-105
 Mechanic arts test, 179
 Median, computation of, 510-513, problems in computation of, 513, summary of steps in computation of, 512-513
 Meier, N C, 443, 456-457, 460, 461
 Meier Art Judgment Test, 455, 456, 457
 Meier-Seashore Art Judgment Test, 456
 Melby, E O, 128, 266, 302, 324, 352, 399, 419, 476, 597
 Mental age, 225-226
 Mental growth, 227-228
 Mental index, 241
Mental Measurements Yearbooks, 111
 Mental test, meaning of, 10-11
 Merrill, M A, 35, 206-209, 222, 226, 232
 Merrill-Palmer Scale of Mental Tests, 232
 Metronoscope, 247
 Metropolitan Achievement Tests, 87, 119, 186, 278-280, 280-283, 481, 486, 488-489, 491
 Metropolitan Readiness Tests, 216-218, 335
 Meyer, G, 141
 Michell, E, 420
 Mid-measure, computation of, 509-510, problems in computation of, 513
 Miller, G S, 420
 Minneapolis Self-Corrective Handwriting Charts, 397
 Modern School Achievement Tests, 181, 427, 481, 490
 Monroe, M, 353
 Monroe, W S, 143, 150, 182, 395-396
 Monroe Reading Aptitude Tests, 335
 Moore, J E, 461
 Moore, M W, 274-275
 Morgan, J J B, 287
 Morrison, W R, 475
 Mort, P R, 129, 287, 324, 353, 400, 420
 Morton, N W, 31, 218-219
 Morton, R L, 324
 Multiple-choice items, 177-182, 193-194, 310-311, 312-313, 412-413, 430-431, 450, 472-473
 Multiplex Quick-Score Grader, 122, 124-125
 Munro, T, 461
 Munsterberg, H, 45, 215
 Murra, W F, 420
 Music appreciation tests, 451-452
 Music education, 444-452; aims of, 444-445, bibliography, 459-461, measurable qualities in, 444, measurement of ability in, 445-452, remediation in, 452
 Musical knowledge tests, 448-450
 Musical memory tests, 447-449
 Musical skills tests, 451
 Musical talent tests, 445-447
 Myers, G C, 324
 Nash, J B, 477
 Nation-wide testing programs, 109
 National Achievement Tests, 185, 312, 410, 431, 465
 National Council of Teachers of English, 111
 Nature study, objectives of, 424
 Neave, E M, 420
 Nelson, N P, 475, 477
 Nelson, M J, 35, 73, 129, 171, 185, 198, 222, 324, 353, 401, 413, 414, 420, 461, 491, 581
 Netzer, R F, 360
 New Revised Stanford-Binet Tests of Intelligence, 206-209, 232
 New Stanford Achievement Test, 427, 484
 New-type test See *Informal Objective Tests*
 Noll, V H, 422-423, 436-437, 441
 Norms, age, 86-87, derivation of, 84-93, grade, 85-86, for interpreting test results, 568-581, in measurement of progress, 588-589, percentile, 88-91, reliability of, 94-95, vs standards, 91-93
 Nystrom, E. C, 397
 Objective test, meaning of, 15-16
 Objective tests, early, 40-41, first in America, 41, informal objective, 25-26, 151-198, standardized, 16-24, 74-97, 99-105, standardized vs informal objective, 153-155
 Objectivity, determination of test, 567, of a test, 66-68, of scoring, 157, of test items, 76-78
 Objectivity coefficient, 67, 567
 Observation methods, 249-250, 439
 Odell, C W, 35, 46, 51, 73, 97, 129, 143-144, 150, 198, 222, 243, 548, 597
 Odell Scales for Rating Pupils' Answers to Nine Types of Thought Questions in General Science, 428
 Ogive curve See *Percentile Curve*
 Ohio State University, 162
 Olson, W C, 245, 259-260, 267, 287

- Ophthalmograph, 347
- Oral examination, advantages of, 132-133; only uses of, 37; limitations of, 132; nature of, 15; uses of, 132-133
- Oral language, diagnosis of ability in, 360-362; scales, 360; skills, 353-354
- Oral questioning, meaning of, 15
- Oral reading, analysis and diagnosis of ability in, 336-339; check tests, 337-339; paragraphs, 336-337; remedial drill in, 347-349; vs. silent reading, 332-337
- Orbita, P. T., 430
- Orleans, J. S., 35, 73, 120, 130, 193, 222, 243, 302, 331, 337, 310
- O'Shea, M. V., 374
- Otis, A. S., 41, 114-115, 210, 231, 326, 343
- Otis Classification Test, 100
- Otis Quick-Scoring Tests of Mental Ability, 211
- Otis Scale for Rating Tests, 114-115
- Overman, J. R., 305-306
- Palmer, G. T., 477
- Paris, University of, 38
- Patterson, D. G., 103, 213, 220
- Paula, E. M., 302
- Pearson, K., 42
- Pearson product-moment correlation coefficient, computation of, 532-537; problem in computation of, 541; summary of steps in computation of, 532-533
- Percentile curve, 545; deviation in constructing, 546
- Percentile norms, 38-41
- Percentile ranks, 231-232, 245-246, 580; problem in computing, 540
- Percentiles, as derived scores, 557-558; problem in finding, 530
- Performance, definition of, 40
- Performance test, meaning of, 30
- Performance tests, 218-220, 238
- Personal Attitudes Test for Younger Boys, 253-254
- Personal constant, 230-231
- Personal report blanks, 251, 261-262
- Personality, 605; measurement of, 247-251, 264-266; nature of, 244-247
- Personality quotient, 264-266
- Personality test, meaning of, 11, 30-31
- Personality tests, development of, 40; of adjustment, 259-263; of attitudes, 253-254; of interests, 255-257; of total personality, 264-266; types of, 30-33
- Peterson, J., 42, 51, 222, 243
- Physical classification tests, 474-475
- Physical education, 468-476, 608; bibliography, 476-478; diagnosis in, 475; measurement of ability in, 469-475; objectives of, 468-470
- Physical examinations, 467
- Physical qualities, tests of general, 470-471
- Physiology, objectives of, 474
- Pintner, R., 28, 35, 51, 200, 218, 220, 222, 241, 243, 262
- Pintner Aspects of Personality, 261
- Pintner Educational Achievement Tests, 400
- Pintner General Ability Tests, 28
- Pintner-Paterson Performance Scale, 218-220
- Piper, A. H., 29
- Plato, 37
- Pond, F. L., 333
- Poppeno, H. E., 429, 432
- Posture tests, 471
- Power test, meaning of, 12-14
- Powers, S. R., 34, 170, 184, 422, 432, 441
- Powers General Science Test, 59
- Preparation, normal objective test, 163-164; standard test, 95-96
- Pressey, L. C., 404
- Pressey, S. L., 21
- Pressey Diagnostic Tests in English Composition, 106-107
- Pressey Interest-Attitude Test, 32-33, 253, 255
- Prevention, in health education, 467-468
- Price, R. A., 420
- Primary mental abilities, 202
- Printing, or test, 95-96
- Problem solving, measurement of ability in, 311-312, 316-318
- Problem-solving tests, in social studies, 407
- Problems, in assigning marks for test scores, 203; in assigning relative ranks, 243; in computing and crashing percentile data, 540; in computing the arithmetic mean, 500; in computing the mid-mean and median, 511; in computing the quartile deviation, 521; in computing the standard deviation, 528; in computing T-scores, 560; in estimating test reliability, 567; in interpreting test scores, 570-581; in tabulating test scores, 502-503; in calculating the correlation coefficient, 541
- Product scale, meaning of, 22-24
- Profile charts, 270-281
- Prognostic test, meaning of, 16-18
- Progress, norms in measurement of, 583-580; records, 278-280
- Progressive Achievement Tests, 23, 100, 276-278, 311, 312, 313
- Progressive Education Association, 48

- Progressive School Achievement Tests, 490
- Projective methods, 248
- "Pseudo-sciences," 245
- Psychological examinations See *Intelligence Tests*
- Public School Achievement Tests, 429, 490
- Pupil diagnosis, 99-100
- Pupil gradation, 100-102
- Pupil guidance, 99
- Pupils, classification and placement of, 585-586, sectioning of, 586-588
- Quality scale, meaning of, 22-23
- Quartile deviation, as measure of variability, 517-518, computation of, 518-521, problems in computation of, 520-521, summary of steps in computation of, 519-520
- Quintilian, 37-38
- Quotient, accomplishment, 239-241, educational, 126-127, intelligence, 226-230, personality, 264-266
- Randall, J. H., 325
- Range, 496, 517
- Ranking scores, need for method of, 540-542
- Ranks, assignment of percentile, 543-546, assignment of relative, 542-543, problems in assignment of percentile, 546, problems in assignment of relative, 543
- Rate test, meaning of, 12-14
- Rathbone, J. L., 477
- Raths, L. E., 198
- Rating scales, 250-251, 259-261, test, 111-115
- Readiness test, meaning of, 29
- Readiness tests, 216-218, in reading, 333-336
- Readings, 326-351, analysis and diagnosis of disability in, 327-331, bibliography, 352-354, corrective exercises in, 347-350, importance of ability in, 326-327, major abilities in, 327-331, measurement of readiness for, 333-336, oral, 336-339, 347-349, oral vs silent, 332-333, silent, 339-347, 349-351
- Reading skills, 328-331
- Reading test, interpretation of results from, 568-570
- Reading tests, 56, 88, 100, 173-174, 176, 180
- Recall items, 26, 161
- Receptive language arts See *Reading*
- Recognition items, 25-26, 161
- Reeve, E. B., 173
- Relationship, need for measures of, 528-529, Pearson product-moment correlation coefficient as measure of, 532-541
- Relative ranks, 542-543, problem in computation of, 543
- Reliability, coefficient, 61-63, 565, of a test, 93-95, 564-567, of norms, 94-95, problems in estimating test, 567
- Remedial instruction, 296-301, in arithmetic, 318-323, in elementary science, 439-440, in handwriting, 394-399, in language, 369-373, in music, 452, in reading, 347-350, in social studies, 418, in spelling, 382-384
- Remmers, H. H., 35, 407
- Remmers Generalized Attitudes Scales, 253, 407
- Research bureaus, 47
- Response, uniformity of, 77
- Retesting coefficient, 62, 565
- Rice, G. A., 198
- Rice, J. M., 41, 50
- Rice, O. S., 350
- Richardson, M. W., 63, 230, 243, 566
- Riemenschneider, A., 461
- Rinsland, H. D., 35, 198, 577, 597
- Rivlin, H. N., 287
- Robertson, D. A., 250
- Rodgers, E. G., 477
- Rogers, C. R., 262
- Rogers, F. R., 470, 477
- Rogers Test of Personality Adjustment, 262
- Ross, C. C., 51, 73, 97, 150, 198, 303, 548, 582
- Ruch, G. M., 20, 35, 48, 51, 73, 97, 129, 136, 154, 171, 173-174, 198, 300, 307-309, 316-317, 320, 322, 411, 420, 429, 430, 432, 450, 451, 491, 575, 582, 593, 597
- Ruch-Popenoe General Science Test, 429, 432
- Rugg, H. O., 597
- Russell, C., 35, 38, 51, 97, 129, 198, 610
- Saetveit, J., 446
- Sampling, extensive, 63-65, 155-157, intensive, 134-135
- Sandiford, P., 222, 233, 243, 267
- Sangren, P. V., 129, 353, 597
- Scale, meaning of, 11-12
- Scale books, 40-41
- Scaled scores, 559
- Scaled test, meaning of, 12
- Scaling of test items, 563-564
- Scheidemann, N. V., 37
- Schindler, A. W., 430
- Schneck, M. R., 221
- Schoen, M., 443, 461

- School surveys, 46-47
 Schrammel, H. E., 88
 Schrammel-Gray High School and College Reading Test, 88
 Science, limitations of measurement in, 425-426
 Science tests, 13, 59, 178, 179, 184, 185
 Scientific attitude, measurement of, 436-438
 Scorability, of a test, 69
 Score card, for rating essay questions, 148
 Scoring, informal objective tests, 165-166; intelligence tests, 223-224, keys, 119-120; objectivity of, 157; subjectivity of, 135-139; tests, 118-125
 Sealy, G. A., 129, 150, 198, 429, 581
 Seashore, C. E., 443, 446, 456, 461
 Seashore Musical Talent Tests, 446
 Seder, M., 129, 287, 548, 582
 Segel, D., 48, 108-109, 287
 Selecting tests, 110-114
 Self-scoring tests, 121-122
 Semi-interquartile range. See *Quartile Deviation*
 Shaffer, L. F., 245, 246, 267, 287
 Shepard, L. A., 368
 Sherman, M., 267, 287
 Sheviakov, G. V., 267
 Shields, G., 350
 Shotwell, A. M., 35, 222, 243, 325, 354, 401, 420, 442, 461, 491
 Shuttleworth, F. K., 267
 Siceloff, L. P., 176
 Sigma. See *Standard Deviation*
 Silance, E. B., 35
 Silent reading, analysis and diagnosis of ability in, 339-347; remedial drill in, 349-351; vs oral reading, 332-333
 Simon, T., 43, 50
 Simple recall items, 170-171, 190-192, 310, 311, 410-411, 428-429, 449
 Sims, V. M., 144-146, 150
 Skeels, H. M., 228
 Skinner, C. E., 288
 Smith, B. O., 73
 Smith, D. V., 401
 Smith, H. L., 73, 222, 243, 325, 353, 401, 420, 441, 582
 Smith, N. B., 353
 Smith, S., 185, 312, 431, 465
 Social studies, 402-419; bibliography, 419-420; diagnosis and remedy in, 418; informal objective tests in, 415-417; measurable qualities in, 406-408; objectives of, 402-403; organization of, 403-404; standardized tests in, 408-415
 Socrates, 37
 Sommer, R., 49
 Sones, W. W. D., 183-184
 Sones-Harry High School Achievement Test, 183-184
 Soper, W., 450
 Sorenson, H., 548
 Souders, L. B., 150
 Spache, G., 401
 Spanev, E., 176
 Spartans, 38
 Spaulding, E. R., 367
 Spearman, C., 45, 202, 222
 Spearman-Brown Prophecy Formula, 63, 565
 Specific intelligence test, meaning of, 28
 Specific intelligence tests, 215-218; of aptitude, 215-216; of readiness, 216-218; types of, 28-29; uses of, 237-238
 Speer, R. K., 184, 312, 431, 465
 Spelling, 373-384; bibliography, 399-401; conscience, 382; consciousness, 382; construction of tests in, 376-379; diagnosis of disability in, 379-382; importance of ability in, 373-374; measurable qualities in, 375; purposes of teaching, 374-375; remediation in, 382-384
 Spelling scale, 24
 Spitzer, H. F., 17, 120, 177, 310
 Sports proficiency tests, 474
 Staffelbach, E. H., 399
 Stagner, R., 35, 267
 Stalnaker, J. M., 145, 146
 Standard deviation, as measure of variability, 521-523; computation of, 523-528; practical uses of, 560-563; problems in computation of, 528; summary of steps in computation of, 527-528
 Standard measures, 558-559
 Standard scores, 232, 559
 Standardization, meaning of, 74-75
 Standardized scales, quality and product, 22-24
 Standardized test, meaning of, 15-16; vs informal objective test, 153-155
 Standardized tests, analytic and diagnostic, 18-22; construction of, 74-97; early development of, 45-47; limitations as measures of improvement, 589; survey and prognostic, 16-18; uses of, 99-105, 601-603
 Standards, vs norms, 91-93
 Stanford Achievement Test, 171, 173-174, 293-295, 367, 427, 430, 480, 481, 483-484, 486, 491
 Stanford-Binet Tests of Intelligence, 206-209, 232
 Stanford Revision of Binet-Simon Scales, 44
 Stanton, H. M., 461
 Starch, D., 133, 135-136, 142
 State-wide testing programs, 108-109

- Statistical method, foundations of, 42, basic procedures in, 493-548
- Stern, W., 200
- Stevenson Problem Analysis Tests, 318
- Stewart, A. W., 442
- Stickney, G. E., 367
- Stoddard, G. D., 97, 129, 222, 228-229, 243, 597
- Stogdill, E. L., 267, 287
- Stone, C. R., 334-335
- Stone, C. W., 45-46
- Stone, M. B., 325
- Stone-Grover Classification Test for Beginners in Reading, 334-335
- Strang, R., 463, 464, 465, 477
- Strecker, E. A., 266
- Studbaker, J. W., 20, 300, 307-309, 316-317, 320, 322, 575
- Stodouress, index of, 241
- Subjectivity of scoring, 135-139
- Sub-total, 511
- Sullivan, E. F., 212
- Superintendence, Department of, 41, 46
- Superstitious beliefs, measurement of, 438-439
- Survey test, meaning of, 16-17
- Symonds, P. M., 35, 73, 166, 197, 222, 228, 241, 243, 267, 287, 548, 582
- T-scores, 559-560, problem on computation of, 560
- Tabulation of test scores in frequency distribution, 494-503, in double-entry table, 532
- Teacher-made test, meaning of, 25
- Teacher-made tests, essay, 25, 133-150, informal objective, 25-26, 151-198, uses of, 600-601
- Teacher's Word Book, The*, 377
- Teachers' marks, 46, functions of, 591-592, objectifying, 592-597, subjectivity of, 590-591
- Teaching method, evaluation of, 588
- Telesync, 346
- Term in, L. M., 35, 44, 171, 173-174, 200, 206-209, 210, 222, 225, 226, 232, 243, 410, 491
- Terman Group Test of Mental Ability, 103
- Test, adequacy, 63-65, administrability, 68-69, comparability, 69-70, economy, 70-71, meaning of, 11, objectivity, 66-68, reliability, 61-63, scorability, 69, utility, 71, validity, 52-61
- Test items, alternate-response, 174-177, completion, 172-174, constructing and validating, 76-81, difficulty of, 78-80, discriminative power of, 80-81, matching, 182-187, multiple-choice, 177-182, objectivity of, 76-78, scaling of, 563-564, simple recall, 170-171, suggestions for construction of, 188-196
- Test rating scales, 111-115
- Test results, interpretation of, 568-581, 603-604
- Test score, meaning of, 549-552
- Test scores, problems in interpreting, 579-581, tabulation of, 494-503
- Testing, meaning of, 67, preparation for, 116
- Testing program, planning of, 105-110
- Tests, classification of, 10-14, disciplinary uses of, 585, educational, 14-26, hand-scored, 118-121, intelligence, 26-30, 205-220, personality, 30-33, 253-266, printing of, 95-96, recognition of need for, 598-599, selection of, 110-114, self-scoring, 121-122, what they do, 4-5, what they do not do, 5, when to give, 107
- Theophrastus, 244-245
- Thorndike, E. L., 45, 46, 51, 200, 228, 325, 353-354, 375, 377, 401, 443
- Thorndike Extension of the Hillebr's English Composition Scale, 365
- Thorndike Scale for Handwriting of Children, 46, 390-391, 394
- Thorpe, L. P., 35, 267, 281, 287
- Thurstone, L. L., 35, 202, 213, 222, 252, 548
- Thurstone, T. G., 213
- Thurstone Scales of Social Attitudes, 253, 407
- Tidman, W. F., 383
- Tiegs, E. W., 97, 100, 129, 198, 212, 281, 303, 311, 312, 313, 325, 354, 401, 420, 441, 477, 548, 582
- Todd, J., 461
- Tool subjects, 605-606
- Torgerson, T. L., 429
- Tormey, T. J., 405
- Tribe, M. R., 365
- Traditional examination See *Essay Examination*
- Triphagen, V., 286, 302
- Travis, L. E., 161-162, 401
- Traxler, A. E., 180, 246, 267, 287, 288, 303, 354, 401
- Traxler Silent Reading Test, 180
- Two-factor theory, 202
- Tyler, R. W., 48, 51, 154-155, 161, 198, 303, 420, 442
- Uhl, W. L., 461
- Underhill, O. E., 179, 184, 432
- Ungrouped data, computation of arithmetic mean from, 503-504, computation of mid-measure from, 509-510, computation of standard deviation from, 523-524

- Unit Scales of Attainment, 13, 173, 367,
430, 480, 481, 485-488, 491
Usage, measurement of ability in, 366-
368
Utility, of a test, 71
- Validity, coefficient, 57-59, curricular,
55-57, of a test, 52-61, 93-94, 564, of
test content, 75-76, psychological and
logical, 60-61, statistical, 57-60
- Van Wageningen, M J, 13, 129, 303, 310,
313, 346
- Van Wageningen General Science Reading
Scales, 428
- Van Wageningen Reading Readiness Tests,
335-336
- Variability, measures of, 515-528, need
for measures of, 515-517, quartile
deviation as measure of, 517-521,
range as measure of, 517, standard
deviation as measure of, 521-528
- Variable errors, 139
- Vocational guidance, 236
- von Borgeersrode, F, 112-114
- Votaw, D F, 100-101, 548
- Walker, H M, 548
- Wallin, J E W, 267
- Walther, E C, 411, 415
- Washburne, C, 325
- Watson, G, 45, 49, 51, 222
- Watson, R E, 430-431
- Wayman, A, 478
- Webb, L W, 35, 222, 243, 325, 354,
401, 420, 442, 461, 491
- Weidemann, C C, 198
- Wellman, B L, 228
- Wesley, E B, 402-403, 416-417, 420
- West, P V, 391
- West, J Y, 439, 442
- Whipple, G M, 44
- Whitford, W G, 443, 454-455, 461
- Whitney, A, 478
- Whittaker, R L, 325
- Wickman, E K, 259-260
- Wiedefeld, N T, 411, 415
- Wiedefeld-Walther Geography Test, 411,
415
- Wilke, W H, 401
- Williams, C L, 325
- Williams, J H, 548
- Willing Scale for Measuring Written
Composition, 366
- Wilson, G M, 325, 354, 401, 420,
478, 597, 610
- Wilson, H E, 420
- Winslow, L L, 453, 460
- Wissler, C, 43
- Witty, P A, 288, 354
- Wood, B D, 97, 154, 288, 412
- Woodrow, H H, 200
- Woodworth, R S, 49
- Woodworth Personal Data Sheet, 49
- Woodv, C, 129, 413, 597
- Woodyard, E, 228
- Worcester, D A, 198
- Workman, L L, 35, 128, 197, 302, 548,
581
- Wrenn, C G, 302
- Wright, W W, 73, 222, 243, 325, 353,
401, 420, 441, 582
- Wrightstone, J W, 146, 354, 420, 442
- Wrightstone Test of Critical Thinking in
the Social Studies, 410
- Written language, skills, 359, 362-365
- Yerkes, R M, 44
- Yokam, G A, 44, 328-331
- Z-scores, 558-559
- Zapf, R M, 438-439, 442
- Zimmerman, J G, 430-431

